3

Reliability

3.1 Classical Test Theory

3.1.1 Overview

To understand the concept of reliability, it is helpful to understand *classical test theory* (CTT), which is also known as "true score theory." Classical test theory is one of various measurement theories, in addition to item response theory and generalizability theory. CTT has been the predominant measurement theory through the history of psychology. CTT is a theory of how test scores relate to a construct. A *construct* is the concept or characteristic that a measure is intended to assess.

Assume you take a measurement of the same object multiple times (i.e., repeated measure). For example, you assess the mass of the same rock multiple times. However, you obtain different estimates of the rock's mass each time. There are multiple possible explanations for this observation. One possible explanation could be that the rock is changing in its mass, which would be consistent with the idea proposed by the Greek philosopher Heraclitus that nothing is stable and the world is in flux. An alternative possibility, however, is that the rock is stable in its mass but the measurements are jittery—that is, they include error.

Based on the possibility that the rock is stable and that differences in scores across time reflect measurement error, CTT proposes the true score formula in Equation (3.1):

$$\begin{gathered} X = T + e \\ \text{observed score} = \text{true score} + \text{measurement error} \end{gathered} \tag{3.1}$$

X is the observed measurement or score, T is the classical (or psychometric) "true score," and e is the measurement error (i.e., error score).

This formula is depicted visually in the form of a path diagram in Figure 3.1.

It is important to distinguish between the classical true score and the Platonic true score. The *Platonic true score* is the truth, and it does not depend on measurement. The Platonic true score is an abstract notion because it is not directly observable and is based on Platonic ideals and theories of the construct and what a person's "true" level is on the construct. In CTT, we attempt to approximate the Platonic true score with the classical true score, (T). If we took infinite repeated observations (and the measurements had no carryover effect), the average score approaches the classical true score, T. That is, $\overline{X} = T$ as number of observations $\rightarrow \infty$. CTT attempts to partition variance into different sources. *Variance* is an index of scores' variability, i.e., the degree to which scores differ. Variance is defined as the average squared deviation from the mean, as in Equation (3.2):

$$\sigma^2_X = E[(X - \mu)^2] \tag{3.2}$$

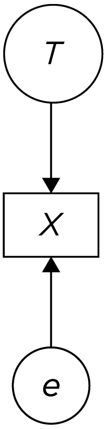


FIGURE 3.1 Classical Test Theory Formula in a Path Diagram.

According to CTT, any observed measurement includes both true score (construct) variance (σ^2_T) and error variance (σ^2_e). Given the true score formula ($X = T + e$), this means that their variance is as follows (see Equation (3.3)):

$$\begin{gathered}\sigma^2_X = \sigma^2_T + \sigma^2_e \\ \text{observed score variance} = \text{true score variance} + \text{error variance}\end{gathered} \tag{3.3}$$

Nevertheless, the *classical true score*, T, is the expected value and is not necessarily the same thing as the Platonic true score (Borsboom, 2003; Klein & Cleary, 1969) because the expected value would need to be entirely valid/accurate (i.e., it would need to be the construct score) for it to be the Platonic true score. The expected score could also be influenced by systematic sources of error such as other constructs, which would not fall into the error portion of the CTT formula because, as described below, CTT assumes that error is random (not systematic). The distinctions between construct score, (classical) true score, and observed score, in addition to validity, reliability, systematic error, and random error are depicted in Figure 3.2.

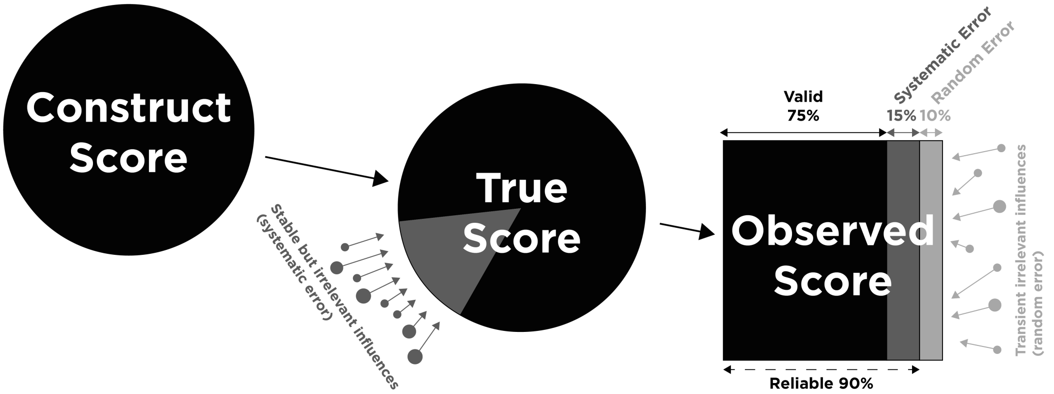

Construct Score
True Score
Stable but irrelevant influences (systematic error)
Valid 75%
Systematic Error 15%
Random Error 10%
Observed Score
Reliable 90%
Transient irrelevant influences (random error)

FIGURE 3.2 Distinctions Between Construct Score, True Score, and Observed Score, in Addition to Reliability, Validity, Systematic Error, and Random Error. (Adapted from W. Joel Schneider.)

The true score formula is theoretically useful, but it is not practically useful because it is an under-identified equation and we do not know the values of T or e based on knowing the observed score (X). For instance, if we obtain an observed score of 10, our formula is $10 = T + e$, and we do not know what the true score or error is. As a result, CTT makes several simplifying assumptions so we can estimate how stable (reliable) or noisy the measure is and what proportion of the observed score reflects true score versus measurement error.

1. $E(e) = 0$

   The first assumption of CTT is that the expected value of the error (i.e., error scores) is zero. Basically, the error component of the observed scores is expected to be random with a mean of zero. The likelihood that the observed score is an overestimate of T is assumed to be the same as the likelihood that the observed score is an underestimate of T. In other words, the distribution of error scores above T is assumed to be the same as the distribution of error scores below T. In reality, this assumption is likely false in many situations. For instance, social desirability bias is a systematic error where people rate themselves as better than they actually are; thus, social desirability results in biasing scores in a particular direction across respondents, so such an error would not be entirely random. But using the assumption that the expected value of e is zero also informs that the expected value of the observed score (X) equals the expected value of the true score (T), as in Equation (3.4):

$$\begin{aligned} E(X) &= E(T + e) \\ &= E(T) + E(e) \\ &= E(T) + 0 \\ &= E(T) \end{aligned} \tag{3.4}$$

2. $r_{T,e} = 0$

   The second assumption of CTT is that the correlation between T and e is zero—that is, people's true scores are uncorrelated with the error around their measurement (i.e., people's error scores). However, this assumption is likely false in many situations. For instance, one can imagine that, on a paper-and-pencil intelligence test, scores may have greater error for respondents with lower true scores and may have less error for respondents with higher true scores.

3. $r_{e_1,e_2} = 0$

   The third assumption of CTT is that the error is uncorrelated across time—that is, people's error scores at time 1 (e_1) are not associated with their error scores at time 2 (e_2). However, this assumption is also likely false in many situations. For example, if some people have a high social desirability bias at time 1, they are likely to also have a high social desirability bias at time 2. That is, the error around measurements of participants is likely to be related across time.

These three assumptions are implicit in the path analytic diagram in Figure 3.3, which depicts the CTT approach to understanding reliability of a measure across two time points.

In path analytic (and structural equation modeling) language, rectangles represent variables we observe, and circles represent latent (i.e., unobserved) variables. The observed scores at time 1 (X_1) and time 2 (X_2) are entities we observe, so they are represented by rectangles.

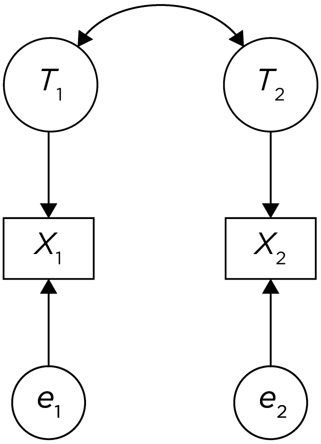


FIGURE 3.3 Reliability of a Measure Across Two Time Points, as Depicted in a Path Diagram.

We do not directly observe the true scores (T_1, T_2) and error scores (e_1, e_2), so they are considered latent entities and are represented by circles. Single-headed arrows indicate regression paths, where conceptually, one variable is thought to influence another variable. As the model depicts, the observed scores are thought to be influenced both by true scores and by error scores. We also expect the true scores at time 1 (T_1) and time 2 (T_2) to be correlated, so we have a covariance path, as indicated by a double-headed arrow. A *covariance* is an unstandardized index of the strength of association between two variables. Because a covariance is unstandardized, its scale depends on the scale of the variables. The covariance between two variables is the average product of their deviations from their respective means, as in Equation (3.5):

$$\sigma_{XY} = E[(X - \mu_X)(X - \mu_Y)] \tag{3.5}$$

The covariance of a variable with itself is equivalent to its variance, as in Equation (3.6):

$$\begin{aligned} \sigma_{XX} &= E[(X - \mu_X)(X - \mu_X)] \\ &= E[(X - \mu_X)^2] \\ &= \sigma^2_X \end{aligned} \tag{3.6}$$

By contrast, a *correlation* is a standardized index of the strength of association between two variables. Because a correlation is standardized (fixed between [-1,1]), its scale does not depend on the scales of the variables. In this figure, no other parameters (regressions, covariances, or means) are specified, so the following are implicit in the diagram:

- $E(e) = 0$
- $r_{T,e} = 0$
- $r_{e_1,e_2} = 0$

The factor loadings reflect the magnitude that the latent factor influences the observed variable. In this case, the true scores influence the observed scores with a magnitude of $\sqrt{r_{xx}}$, which is known as the *index of reliability*. The index of reliability is the theoretical estimate of the correlation between the true scores and the observed scores. This is depicted in Figure 3.4.

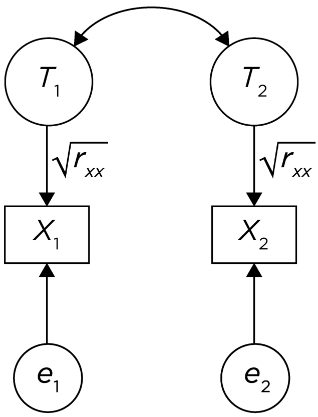


FIGURE 3.4 Reliability of a Measure Across Two Time Points, as Depicted in a Path Diagram; Includes the Index of Reliability.

We can use path tracing rules to estimate the reliability of the measure, where the reliability of the measure, i.e., the *coefficient of reliability* (r_{xx}), is estimated as the correlation between the observed score at time 1 (x_1) and the observed score at time 2 (x_2). According to path tracing rules (Pearl, 2013), the correlation between x_1 and x_2 is equal to the sum of the standardized coefficients of all the routes through which x_1 and x_2 are connected. The contribution of a given route to the correlation between x_1 and x_2 is equal to the product of all standardized coefficients on that route that link x_1 and x_2 that move in the following directions: (a) forward (e.g., T_1 to x_1) or (b) backward once and then forward (e.g., x_1 to T_1 to T_2 to x_2). Path tracing does not allow moving forward and then backward—that is, it does not allow retracing (e.g., e to x_1 to T_1) in the same route. It also does not allow passing more than one curved arrow (covariance path) or through the same variable twice in the same route. Once you know the contribution of each route to the correlation, you can calculate the total correlation between the two variables as the sum of the contribution of each route. Therefore, using one route, we can calculate the association between x_1 and x_2 as in Equation (3.7):

$$\begin{aligned} r_{x_1,x_2} &= \sqrt{r_{xx}} \times r_{T_1,T_2} \times \sqrt{r_{xx}} \\ &= r_{T_1,T_2} \times r_{xx} \\ &= \text{correlation of true scores across time} \times \text{reliability} \end{aligned} \tag{3.7}$$

When dealing with a stable construct, we would assume that the correlation between true scores across time is 1.0: $r_{T_1,T_2} = 1.0$, as depicted in Figure 3.5.

Then, to calculate the association between x_1 and x_2 of a stable construct, we can use path tracing rules as in Equation (3.8):

$$\begin{aligned} r_{x_1,x_2} &= \sqrt{r_{xx}} \times r_{T_1,T_2} \times \sqrt{r_{xx}} \\ &= \sqrt{r_{xx}} \times 1 \times \sqrt{r_{xx}} \\ &= r_{xx} \\ &= \text{coefficient of reliability} \end{aligned} \tag{3.8}$$

That is, for a stable construct (i.e., whose true scores are perfectly correlated across time; $r_{T_1,T_2} = 1.0$), we estimate reliability as the correlation between the observed scores at time

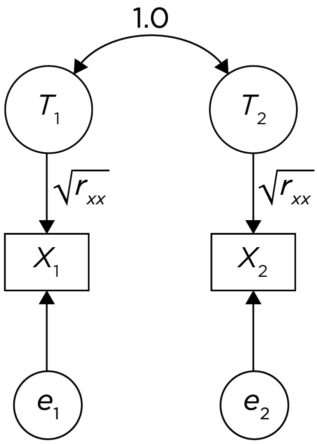


FIGURE 3.5 Reliability of a Measure of a Stable Construct Across Two Time Points, as Depicted in a Path Diagram.

1 (x_1) and the observed scores at time 2 (x_2). This is known as *test–retest reliability*. We therefore assume that the extent to which the correlation between x_1 and x_2 is less than one reflects measurement error (an unstable measure), rather than people's changes in their true score on the construct (an unstable construct).

As described above, the reliability coefficient (r_{xx}) is the association between a measure and itself over time or with another measure in the domain. By contrast, the *reliability index* (r_{xT}) is the correlation between observed scores on a measure and the true scores (Nunnally & Bernstein, 1994). The reliability index is the square root of the reliability coefficient, as in Equation (3.9).

$$\begin{aligned} r_{xT} &= \\ &= \sqrt{r_{xx}} \\ &= \text{index of reliability} \end{aligned} \tag{3.9}$$

3.1.2 Four CTT Measurement Models

There are four primary measurement models in CTT (J. M. Graham, 2006):

1. parallel
2. tau-equivalent
3. essentially tau-equivalent
4. congeneric

3.1.2.1 Parallel

The parallel measurement model is the most stringent measurement model for use in estimating reliability. In CTT, a measure is considered parallel if the true scores and error scores are equal across items. That is, the items must be unidimensional and assess the same construct, on the same scale, with the same degree of precision, and with the same amount of error (J. M. Graham, 2006). Items are expected to have the same strength of association (i.e., factor loading or discrimination) with the construct.

3.1.2.2 Tau-Equivalent

The tau (τ)-equivalent measurement model is the same as the parallel measurement model, except error scores are allowed to differ across items. That is, a measure is tau-equivalent if the items are unidimensional and assess the same construct, on the same scale, with the same degree of precision, but with possibly different amounts of error (J. M. Graham, 2006). In other words, true scores are equal across items but each item is allowed to have unique error scores. Items are expected to have the same strength of association with the construct. Variance that is unique to a specific item is assumed to be error variance.

3.1.2.3 Essentially Tau-Equivalent

The essentially tau (τ)-equivalent model is the same as the tau-equivalent measurement model, except items are allowed to differ in their precision. That is, a measure is essentially tau-equivalent if the items are unidimensional and assess the same construct, on the same scale, but with possibly different degrees of precision, and with possibly different amounts of error (J. M. Graham, 2006). The essentially tau-equivalent model allows item true scores to differ by a constant that is unique to each pair of variables. The magnitude of the constant reflects the degree of imprecision and influences the mean of the item scores but not its variance or covariances with other items. Items are expected to have the same strength of association with the construct.

3.1.2.4 Congeneric

The congeneric measurement model is the least restrictive measurement model. The congeneric measurement model is the same as the essentially tau (τ)-equivalent model, except items are allowed to differ in their scale. That is, a measure is congeneric if the items are unidimensional and assess the same construct, but possibly on a different scale, with possibly different degrees of precision, and with possibly different amounts of error (J. M. Graham, 2006). Items are not expected to have the same strength of association with the construct.

3.2 Measurement Error

Measurement error is the difference between the measured (observed) value and the true value. All measurements come with uncertainty and measurement error. Even a measure of something as simple as whether someone is dead has error.

There are two main types of measurement error: systematic (nonrandom) error and unsystematic (random) error. In addition, measurement error can be within-person, between-person, or both.

3.2.1 Systematic (Nonrandom) Error

An example of *systematic error* is depicted in Figure 3.6. Systematic error is error that influences consistently for a person or across the sample. An error is systematic if the error always occurs, with the same value, when using the measure in the same way and in the same case. An example of a systematic error is a measure that consistently assesses constructs other than the construct the measure was designed to assess. For instance, if a test written in English to assess math skills is administered in a nonnative English-speaking country, some portion of the scores will reflect variance attributable to English reading skills rather than the construct of interest (math skills). Other examples of systematic error include response styles or subjective, idiosyncratic judgments by a rater—for instance, if the rater's judgments are systematically harsh or lenient. A systematic error affects the average score (i.e., resulting in bias), which makes the group-level estimates less accurate and makes the measurements for an individual less accurate.

As depicted in Figure 3.6, systematic error does not affect the variability of the scores, but it does affect the mean of the scores, so the person-level mean and group-level mean are less accurate. In other words, a systematic error leads to a biased estimate of the average. However, multiple systematic errors may simultaneously coexist and can operate in the same direction (exacerbating the effects of bias) or in opposite directions (hiding the extent of bias).

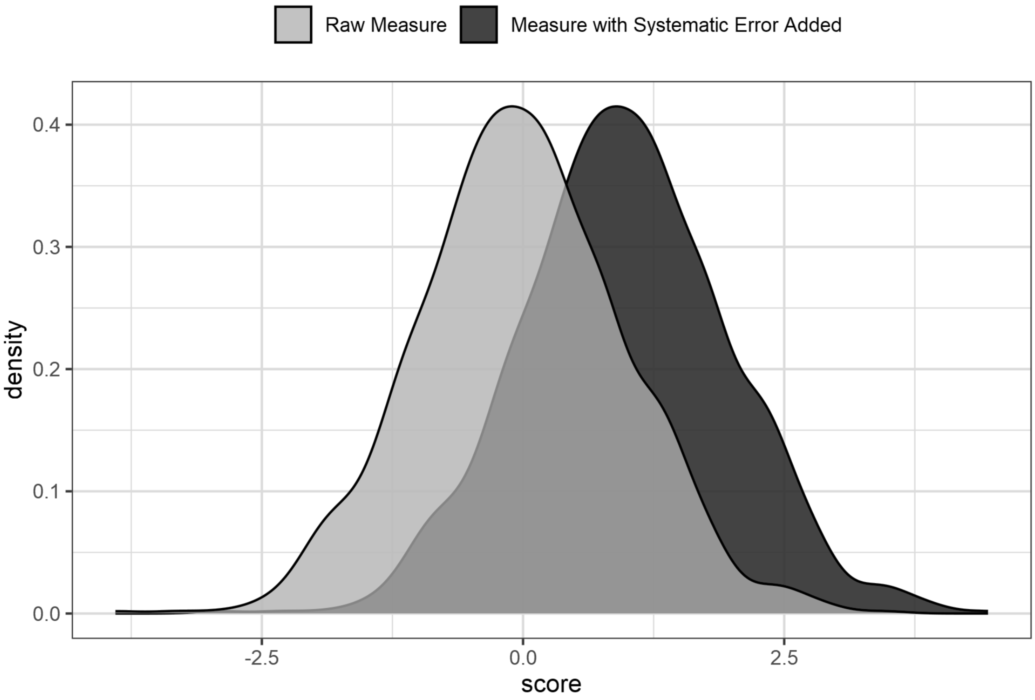


FIGURE 3.6 Systematic Error.

3.2.2 Unsystematic (Random) Error

An example of *unsystematic (random) error* is depicted in Figure 3.7. Random error occurs due to chance. For instance, a random error could arise from a participant being fatigued on a particular testing day or from a participant getting lucky in guessing the correct answer. Random error does not have consistent effects for a person or across the sample, and it may vary from one observation to another. Random error does not (systematically) affect the average, i.e., the group-level estimate—random error affects only the variability around the average (noise). However, random error makes measurements for an individual less accurate. A large number of observations of the same construct cancels out random error but does not cancel out systematic error.

As depicted in Figure 3.7, random error does not affect the mean of the scores, but it does increase the variability of the scores. In other words, the group-level mean is still accurate, but individuals' scores are less precise.

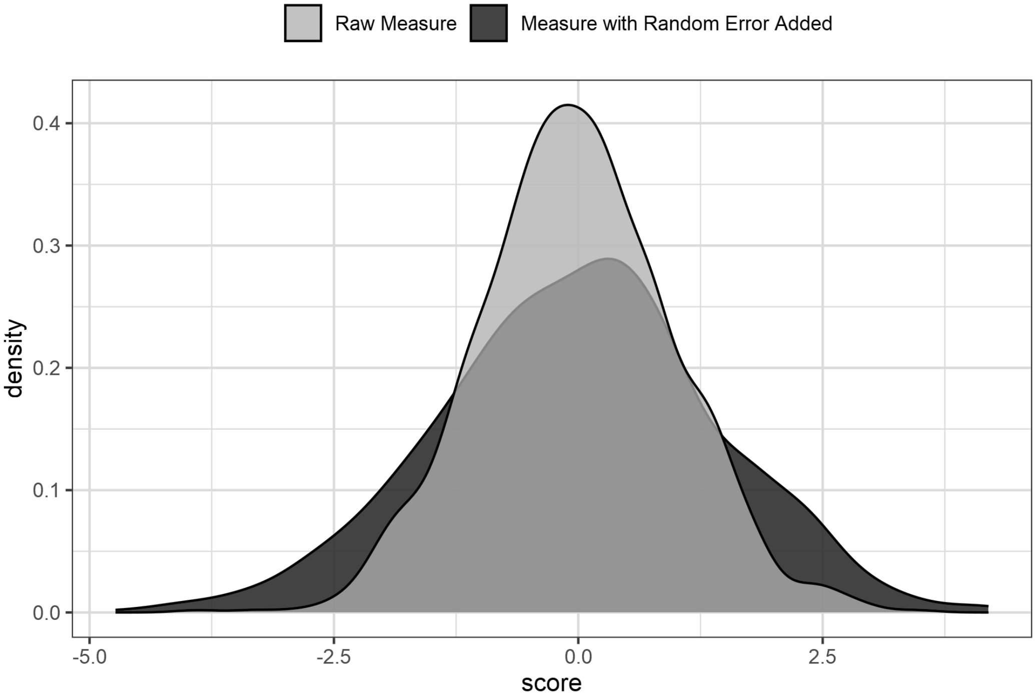


FIGURE 3.7 Random Error.

3.2.3 Within-Person Error

Consider two data columns, one column for participants' scores at time 1 and another column for participants' scores at time 2. Adding within-person error would mean adding noise (e) within the given row (or rows) for the relevant participant(s). Adding between-person error would mean adding noise (e) across the rows within the column.

Within-person error occurs within a particular person. For instance, you could add within-person error to a data set by adding error to the given row (or rows) for the relevant participant(s).

3.2.4 Between-Person Error

Between-person error occurs across the sample. You could add between-person random error to a variable by adding error across the rows, within a column.

3.2.5 Types of Measurement Error

There are four nonmutually exclusive types of measurement error: within-person random error, within-person systematic error, between-person random error, and between-person systematic error. The four types of measurement error are depicted in Figure 3.8, as adapted from Willett (2012).

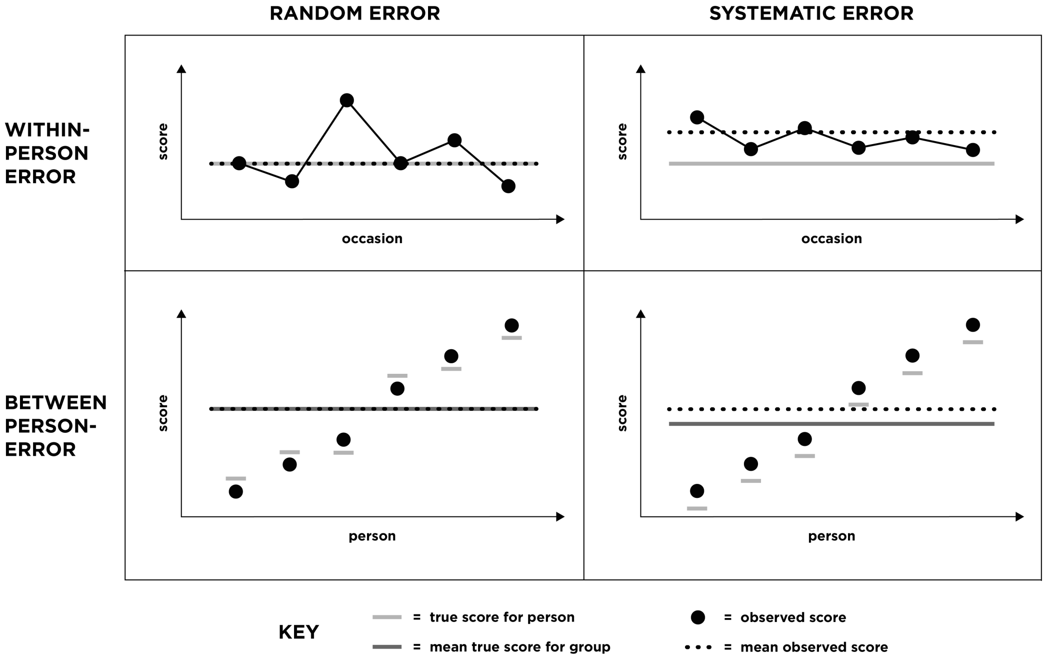


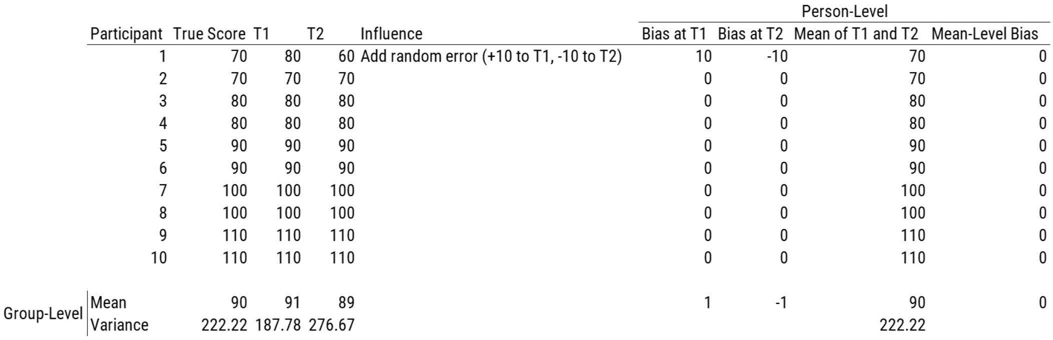
FIGURE 3.8 Types of Measurement Error.

3.2.5.1 Within-Person Random Error

Adding within-person random error would involve adding random noise (e) to the given row (or rows) for the relevant participant(s). This could reflect momentary fluctuations in the assessment for a specific person. When adding within-person random error, the person's and group's measurements show no bias, i.e., there is no consistent increase or decrease in the scores from time 1 to time 2 (at least with a sample size large enough to cancel out the random error, according to the law of large numbers). A person's average approximates their true score if many repeated measurements are taken. A group's average approximates the sample mean's true score, especially when averaging the repeated measures across time. The influence of within-person random error is depicted in Figure 3.9.

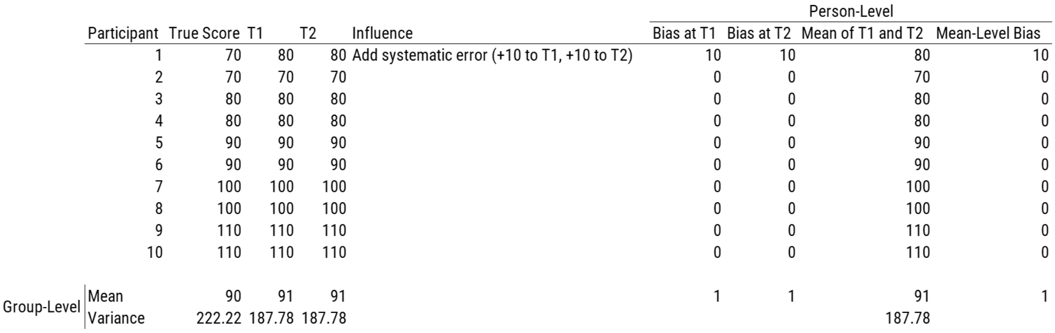
3.2.5.2 Within-Person Systematic Error

Adding within-person systematic error would involve adding systematic noise (e) (the same variance across columns) to the given row (or rows), reflecting the relevant participant(s). These are within-person effects that are consistent across time. For example, social desirability bias is high for some people and low for others. Another instance in which within-person systematic error could exist is when one or more people consistently misinterpret a particular question. Within-person systematic error increases person-level bias because the person's mean shows a greater difference from their true score. The influence of within-person systematic error is depicted in Figure 3.10.



						Person-Level			
	Participant	True Score	T1	T2	Influence	Bias at T1	Bias at T2	Mean of T1 and T2	Mean-Level Bias
	1	70	80	60	Add random error (+10 to T1, -10 to T2)	10	-10	70	0
	2	70	70	70		0	0	70	0
	3	80	80	80		0	0	80	0
	4	80	80	80		0	0	80	0
	5	90	90	90		0	0	90	0
	6	90	90	90		0	0	90	0
	7	100	100	100		0	0	100	0
	8	100	100	100		0	0	100	0
	9	110	110	110		0	0	110	0
	10	110	110	110		0	0	110	0
Group-Level	Mean	90	91	89		1	-1	90	0
	Variance	222.22	187.78	276.67				222.22	

FIGURE 3.9 Within-Person Random Error.



						Person-Level			
	Participant	True Score	T1	T2	Influence	Bias at T1	Bias at T2	Mean of T1 and T2	Mean-Level Bias
	1	70	80	80	Add systematic error (+10 to T1, +10 to T2)	10	10	80	10
	2	70	70	70		0	0	70	0
	3	80	80	80		0	0	80	0
	4	80	80	80		0	0	80	0
	5	90	90	90		0	0	90	0
	6	90	90	90		0	0	90	0
	7	100	100	100		0	0	100	0
	8	100	100	100		0	0	100	0
	9	110	110	110		0	0	110	0
	10	110	110	110		0	0	110	0
Group-Level	Mean	90	91	91		1	1	91	1
	Variance	222.22	187.78	187.78				187.78	

FIGURE 3.10 Within-Person Systematic Error.

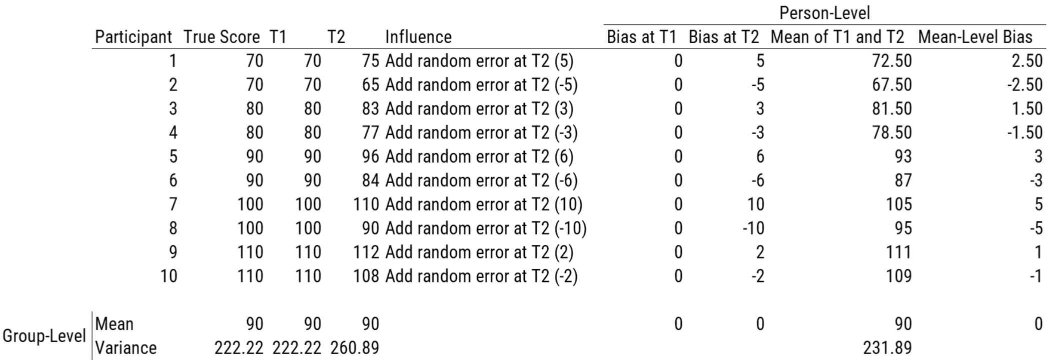
3.2.5.3 Between-Person Random Error

Adding between-person random error at time 2 would involve adding random noise (e) across the rows, within the column.

Between-person random error would result in less accurate scores at the person level but would not result in bias at the group level. At a given timepoint, it results in overestimates of the person's true score for some people and underestimates for other people, i.e., there is no consistent pattern across the sample. Thus, the group average approximates the sample's mean true score (at least with a sample size large enough to cancel out the random error, according to the law of large numbers). In addition, the average of repeated measurements of the person's score would approximate the person's true score. However, the group's variance is inflated. The influence of between-person random error is depicted in Figure 3.11.

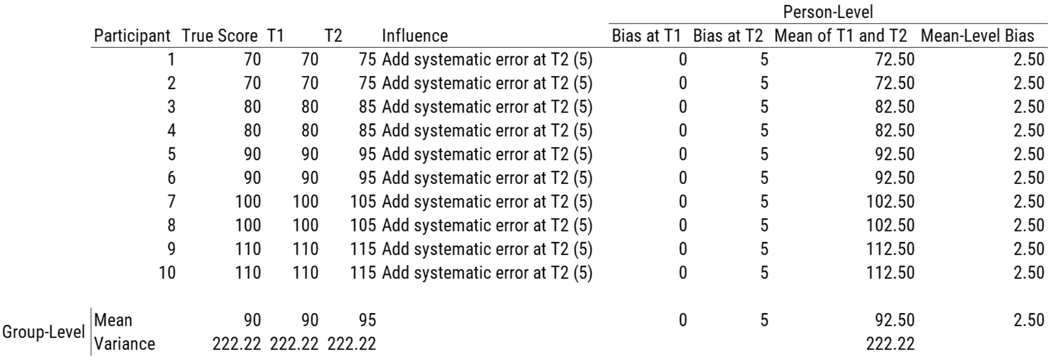
3.2.5.4 Between-Person Systematic Error

Adding between-person systematic error at time 2 would involve adding systematic noise (e) across the rows, within the column. Between-person systematic error results from within-person error that tends to be negative or positive across participants. For instance, this could reflect an influence with a shared effect across subjects. For example, social desirability leads to a positive group-level bias for rating their socially desirable attributes. Another example would be when a research assistant enters values wrong at time 2 (e.g., adding 10 to all participants' scores). Between-person systematic error increases bias because it results in a greater group mean difference from the group's mean true score. The influence of between-person systematic error is depicted in Figure 3.12.



						Person-Level			
	Participant	True Score	T1	T2	Influence	Bias at T1	Bias at T2	Mean of T1 and T2	Mean-Level Bias
	1	70	70	75	Add random error at T2 (5)	0	5	72.50	2.50
	2	70	70	65	Add random error at T2 (-5)	0	-5	67.50	-2.50
	3	80	80	83	Add random error at T2 (3)	0	3	81.50	1.50
	4	80	80	77	Add random error at T2 (-3)	0	-3	78.50	-1.50
	5	90	90	96	Add random error at T2 (6)	0	6	93	3
	6	90	90	84	Add random error at T2 (-6)	0	-6	87	-3
	7	100	100	110	Add random error at T2 (10)	0	10	105	5
	8	100	100	90	Add random error at T2 (-10)	0	-10	95	-5
	9	110	110	112	Add random error at T2 (2)	0	2	111	1
	10	110	110	108	Add random error at T2 (-2)	0	-2	109	-1
Group-Level	Mean	90	90	90		0	0	90	0
	Variance	222.22	222.22	260.89				231.89	

FIGURE 3.11 Between-Person Random Error.



						Person-Level			
	Participant	True Score	T1	T2	Influence	Bias at T1	Bias at T2	Mean of T1 and T2	Mean-Level Bias
	1	70	70	75	Add systematic error at T2 (5)	0	5	72.50	2.50
	2	70	70	75	Add systematic error at T2 (5)	0	5	72.50	2.50
	3	80	80	85	Add systematic error at T2 (5)	0	5	82.50	2.50
	4	80	80	85	Add systematic error at T2 (5)	0	5	82.50	2.50
	5	90	90	95	Add systematic error at T2 (5)	0	5	92.50	2.50
	6	90	90	95	Add systematic error at T2 (5)	0	5	92.50	2.50
	7	100	100	105	Add systematic error at T2 (5)	0	5	102.50	2.50
	8	100	100	105	Add systematic error at T2 (5)	0	5	102.50	2.50
	9	110	110	115	Add systematic error at T2 (5)	0	5	112.50	2.50
	10	110	110	115	Add systematic error at T2 (5)	0	5	112.50	2.50
Group-Level	Mean	90	90	95		0	5	92.50	2.50
	Variance	222.22	222.22	222.22				222.22	

FIGURE 3.12 Between-Person Systematic Error.

3.2.6 Summary

In sum, all types of measurement error (whether systematic or random) lead to less accurate scores for an individual. But different kinds of error have different implications. Systematic and random error have different effects on accuracy at the group-level. Systematic error leads to less accurate estimates at the group-level, whereas random error does not.

CTT assumes that all error is random. According to CTT, as the number of measurements approaches infinity, the mean of the measurements gets closer to the true score, because the random errors cancel each other out. With more measurements, we reduce our uncertainty and increase our precision. According to CTT, if we take many measurements and the average of the measurements is 10, we have some confidence that the true score $(T) \approx 10$. In reality, however, error for a given measure likely includes both systematic and random error.

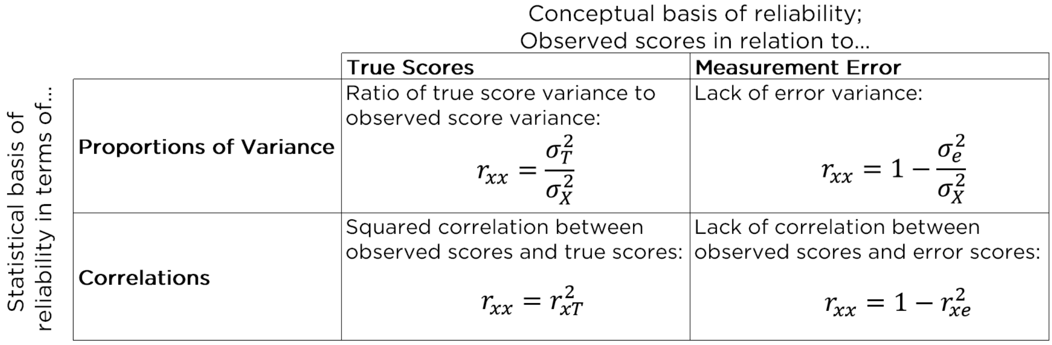
3.3 Overview of Reliability

The "Standards for Educational and Psychological Testing" set the standard for educational and psychological assessment and are jointly published by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. According to the "Standards" (American Educational Research Association et al., 2014, p. 33), reliability is the “consistency of the scores across instances of the testing procedure.” In this book, we define reliability as how much repeatability, consistency, and precision the scores from a measure have.

Reliability (r_{xx}) has been defined mathematically as the proportion of observed score variance (σ^2_X) that is attributable to true score variance (σ^2_T), as in Equation (3.10):

$$\begin{aligned} r_{xx} &= \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_e} \\ &= \frac{\sigma^2_T}{\sigma^2_X} \\ &= \frac{\text{true score variance}}{\text{observed score variance}} \end{aligned} \tag{3.10}$$

An alternative formulation is that reliability (r_{xx}) is the lack of error variance or the degree to which observed scores are correlated with true scores or uncorrelated with error scores. In CTT, reliability can be conceptualized in four primary ways, as depicted in Figure 3.13 (Furr, 2017).



Conceptual basis of reliability;
Observed scores in relation to...

		True Scores	Measurement Error
Statistical basis of reliability in terms of...	Proportions of Variance	Ratio of true score variance to observed score variance: $r_{xx} = \frac{\sigma^2_T}{\sigma^2_X}$	Lack of error variance: $r_{xx} = 1 - \frac{\sigma^2_e}{\sigma^2_X}$
	Correlations	Squared correlation between observed scores and true scores: $r_{xx} = r^2_{xT}$	Lack of correlation between observed scores and error scores: $r_{xx} = 1 - r^2_{xe}$

FIGURE 3.13 Four Different Ways of Conceptualizing Reliability.

However, we cannot *calculate* reliability because we cannot measure the true score component of an observation. Therefore, we *estimate* reliability (the coefficient of reliability) based on the relation between two observations of the same measure (for test–retest reliability) or using other various estimates of reliability.

The coefficient of reliability can depend on several factors. Reliability is inversely related to the amount of measurement error. The coefficient of reliability, like correlation, depends on the degree of spread (variability) of the scores. If the scores at one or both time points show restricted range, the scores will show a weaker association and coefficient of reliability, as shown in Figure 3.19. An ideal way to estimate the reliability of a measure would be to take a person and repeatedly measure them many times to get an estimate of their true score and each measurement’s deviation from their true score, and to do this for many people. However, this is rarely done in practice. Instead of taking one person and repeatedly measuring them many times, we typically estimate reliability by taking many people and doing repeated measures twice. This is a shortcut to estimate reliability, but even this shorter approach is not often done. In short, researchers rarely estimate the test–retest reliability of the measures they use.

Reliability can also be related to the number of items in the measure. In general, the greater the number of items, the more reliable the measure (assuming the items assess the same construct) because we are averaging out random error.

We never see the true scores or the error scores, so we cannot compute reliability—we can only *estimate* it from the observed scores. This estimate of reliability gives a probabilistic answer of the reliability of the measure, rather than an absolute answer.

3.3.1 How Reliable Is Reliable Enough?

As described by Nunnally & Bernstein (1994), how reliable a measure should be depends on the proposed uses. If it is early in the research process, and the focus is on group-level inferences (e.g., associations or group differences), modest reliability (e.g., .70) may be sufficient and save time and money. Then, the researcher can see what the associations would be when disattenuated for unreliability, as described in Section 4.6 of the chapter on validity. If the disattenuated associations are promising, it may be worth increasing the reliability of the measure. Associations are only weakly attenuated above a reliability of .80, so achieving a reliability coefficient of .80 may be an appropriate target for basic research.

However, when making decisions about individual people from their score on a measure, reliability and precision are more important (than when making group-level inferences) because small differences in scores can lead to different decisions. Nunnally & Bernstein (1994) recommend that measures have at least a reliability of .90 and—when making important decisions about individual people—that measures preferably have a reliability of .95 or higher. Nevertheless, they also note that one should not switch to a less valid measure merely because it is more reliable.

3.3.2 Standard Error of Measurement

The estimate of reliability gives a general idea of the degree of uncertainty you have of a person's true score given their observed score. From this, we can estimate the standard error of measurement, which estimates the extent to which an observed score deviates from a true score. The standard error of measurement indicates the typical distance of the observed score from the true score. The formula for the standard error of measurement is in Equation (3.11):

$$\text{standard error of measurement (SEM)} = \sigma_x \sqrt{1 - r_{xx}} \tag{3.11}$$

where σ_x represents the standard deviation of scores. Thus, the standard error of measurement is directly related to the reliability of the measure. The higher the reliability, the lower the standard error of measurement. The standard error of measurement as a function of reliability of the measure and the standard deviation of scores is depicted in Figure 3.14.

The derivation of the SEM (from W. Joel Schneider) is in Equation (3.12):

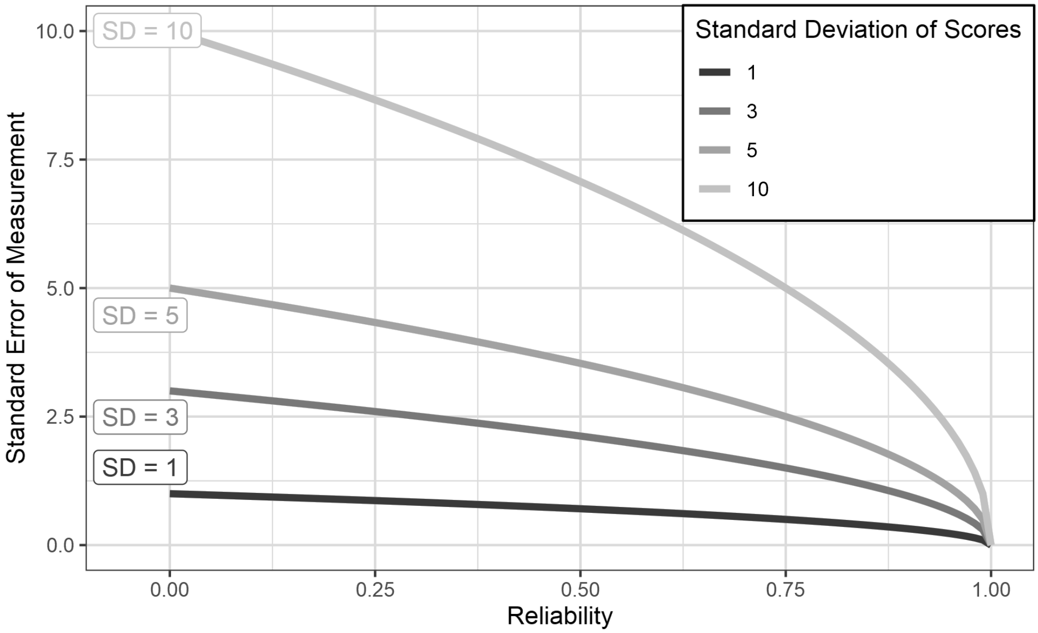


FIGURE 3.14 Standard Error of Measurement as a Function of Reliability.

$$
\begin{aligned}
\text{Remember, based on } X = T + e: && \sigma^2_X &= \sigma^2_T + \sigma^2_e \\
\text{Solve for } \sigma^2_T: && \sigma^2_T &= \sigma^2_X - \sigma^2_e \\
\text{Remember:} && r_{xx} &= \frac{\sigma^2_T}{\sigma^2_X} \\
\text{Substitute for } \sigma^2_T: && &= \frac{\sigma^2_X - \sigma^2_e}{\sigma^2_X} \\
\text{Multiply by } \sigma^2_X: && \sigma^2_X \cdot r_{xx} &= \sigma^2_X - \sigma^2_e \\
\text{Solve for } \sigma^2_e: && \sigma^2_e &= \sigma^2_X - \sigma^2_X \cdot r_{xx} \\
\text{Factor out } \sigma^2_X: && \sigma^2_e &= \sigma^2_X (1 - r_{xx}) \\
\text{Take the square root:} && \sigma_e &= \sigma_X \sqrt{1 - r_{xx}}
\end{aligned}
\tag{3.12}
$$

The SEM is equivalent to the standard deviation of measurement error (e) (Lek & Van De Schoot, 2018), as in Equation (3.13):

$$
\begin{aligned}
\text{standard error of measurement (SEM)} &= \sigma_x \sqrt{1 - r_{xx}} \\
&= \sqrt{\sigma^2_x}\sqrt{1 - r_{xx}} \\
&= \sqrt{\sigma^2_x (1 - r_{xx})} \\
&= \sqrt{\sigma^2_x - \sigma^2_x \cdot r_{xx}} \\
&= \sqrt{\sigma^2_x - \sigma^2_x \frac{\sigma^2_T}{\sigma^2_x}} \\
&= \sqrt{\sigma^2_x - \sigma^2_T} \\
&= \sqrt{\sigma^2_e} \\
&= \sigma_e
\end{aligned}
\tag{3.13}
$$

Around 95% of scores would be expected to fall within ±2 SEMs of the true score (or, more precisely, within ±1.96 SEMs of the true score). In other words, 95% of the time, the true score is expected to fall within ±2 SEMs of the observed score. Given an observed score of $X = 15$ and SEM = 2, the 95% confidence interval of the true score is [11, 19]. So if a person gets a score of 15 on the measure, 95% of the time, their true score is expected to fall within 11–19. An empirical example of estimating the SEM is provided in Section 3.7.

Based on the preceding discussion, consider the characteristics of measures that make them more useful from a reliability perspective. A useful measure would show wide variation across people (individual differences), so we can more accurately estimate its reliability. And we would expect a useful measure to show consistency, stability, precision, and reliability of scores.

3.4 Getting Started

3.4.1 Load Libraries

```
library("petersenlab")
library("psych")
library("blandr")
library("MBESS")
library("lavaan")
library("semTools")
library("psychmeta")
library("irrCAC")
library("gtheory")
library("ggrepel")
library("performance")
library("MOTE")
library("here")
library("tidyverse")
library("tinytex")
library("knitr")
library("kableExtra")
library("rmarkdown")
library("bookdown")
```

3.4.2 Prepare Data

3.4.2.1 Simulate Data

```
sampleSize <- 100

set.seed(52242)

rater1continuous <- rnorm(n = sampleSize, mean = 50, sd = 10)
rater2continuous <- rater1continuous +
  rnorm(
    n = sampleSize,
    mean = 0,
    sd = 4)
rater3continuous <- rater2continuous +
  rnorm(
    n = sampleSize,
    mean = 0,
    sd = 8)

rater1categorical <- sample(
  c(0,1),
  size = sampleSize,
  replace = TRUE)
rater2categorical <- rater1categorical
rater3categorical <- rater1categorical

rater2categorical[
  sample(1:length(rater2categorical),
         size = 10,
         replace = FALSE)] <- 0
rater3categorical[
  sample(1:length(rater3categorical),
         size = 10,
         replace = FALSE)] <- 1

time1 <- rnorm(n = sampleSize, mean = 50, sd = 10)
time2 <- time1 + rnorm(n = sampleSize, mean = 0, sd = 4)
time3 <- time2 + rnorm(n = sampleSize, mean = 0, sd = 8)

item1 <- rnorm(n = sampleSize, mean = 50, sd = 10)
item2 <- item1 + rnorm(n = sampleSize, mean = 0, sd = 4)
item3 <- item2 + rnorm(n = sampleSize, mean = 0, sd = 8)
item4 <- item3 + rnorm(n = sampleSize, mean = 0, sd = 12)
```

```
Person <- as.factor(rep(1:6, each = 8))
Occasion <- Rater <-
  as.factor(rep(1:2, each = 4, times = 6))
Item <- as.factor(rep(1:4, times = 12))
Score <- c(
  9,9,7,4,9,8,5,5,9,8,4,6,
  6,5,3,3,8,8,6,2,8,7,3,2,
  9,8,6,3,9,6,6,2,10,9,8,7,
  8,8,9,7,6,4,5,1,3,2,3,2)
```

3.4.2.2 Add Missing Data

Adding missing data to dataframes helps make examples more realistic to real-life data and helps you get in the habit of programming to account for missing data.

```
rater1continuous[c(5,10)] <- NA
rater2continuous[c(10,15)] <- NA
rater3continuous[c(10)] <- NA

rater1categorical[c(5,10)] <- NA
rater2categorical[c(10,15)] <- NA
rater3categorical[c(10)] <- NA

time1[c(5,10)] <- NA
time2[c(10,15)] <- NA
time3[c(10)] <- NA

item1[c(5,10)] <- NA
item2[c(10,15)] <- NA
item3[c(10)] <- NA
item4[c(10)] <- NA
```

3.4.2.3 Combine Data into Dataframe

```
mydata <- data.frame(
  rater1continuous, rater2continuous, rater3continuous,
  rater1categorical, rater2categorical, rater3categorical,
  time1, time2, time3,
  item1, item2, item3, item4)

pio_cross_dat <- data.frame(Person, Item, Score, Occasion)
```

3.5 Types of Reliability

Reliability is not one thing. There are several types of reliability. In this book, we focus on test–retest, inter-rater, intra-rater, parallel-forms, and internal consistency reliability.

3.5.1 Test–Retest Reliability

Test–retest reliability is defined as the consistency of scores across time. Typically, this is based on a two-week retest interval. The intent of a two-week interval between the original testing and the retest is to provide adequate time to pass to reduce any carryover effects from the original testing while not allowing too much time to pass such that the person's level on the construct (i.e., true scores) would change. A carryover effect is an effect of the experimental condition that affects the participant's behavior at a later time. Examples of carryover effects resulting from repeated measurement can include fatigue, boredom, learning (practice effects), etc. Another potential issue is that measurement error can be correlated across the two measurements.

Test–retest reliability controls for transient error and random response error. If the construct is not stable across time (i.e., people's true scores change), test–retest reliability is not relevant because the CTT approach to estimating reliability assumes that the true scores are perfectly correlated across time (see Section 3.1).

The length of the optimal retest interval depends on the construct of interest. For a construct in which people's levels change rapidly, a shorter retest interval may be appropriate. But one should pay attention to ways to reduce potential carryover effects. By contrast, if the retest interval is too long, people's levels on the construct may change during that span. If people's levels on the construct change from test to retest, we can no longer assume that the true scores are perfectly correlated across time, which would violate CTT assumptions for estimating test–retest reliability of a measure. The longer the retest interval, the smaller the observed association between scores across time will tend to be. For weak associations obtained from a lengthy retest interval, it can be difficult to determine how much of this weak association reflects measurement unreliability versus people's change in their levels on the construct. Thus, when conducting studies to evaluate test–retest reliability, it is important to consider the length of the retest interval and ways to reduce carryover effects.

3.5.1.1 Coefficient of Stability (and Coefficient of Dependability)

The coefficient of stability is the most widely used index when reporting the test–retest reliability of a measure. It is estimated using a Pearson correlation of the scores at time 1 with the score at time 2. That is, the coefficient of stability assesses the stability of individual differences (i.e., rank-order stability). The Pearson correlation is called the coefficient of stability when the length of the retest interval (the delay between test and retest) is on the order of days or weeks. If the retest occurs almost at the same time as the original test (e.g., a 45-minute delay), the Pearson correlation is called the *coefficient of dependability* (Revelle & Condon, 2019).

We estimate the coefficient of stability below:

```
cor.test(x = mydata$time1, y = mydata$time2)
```

```
    Pearson's product-moment correlation

data:  mydata$time1 and mydata$time2
t = 25, df = 95, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8994 0.9539
sample estimates:
   cor
0.9317
```

```
cor(mydata[,c("time1","time2","time3")],
    use = "pairwise.complete.obs")
```

```
       time1  time2  time3
time1 1.0000 0.9317 0.8291
time2 0.9317 1.0000 0.8426
time3 0.8291 0.8426 1.0000
```

The r value of .93 indicates a strong positive association. Figure 3.15 depicts a scatterplot of the time 1 scores on the x-axis and time 2 scores on the y-axis.

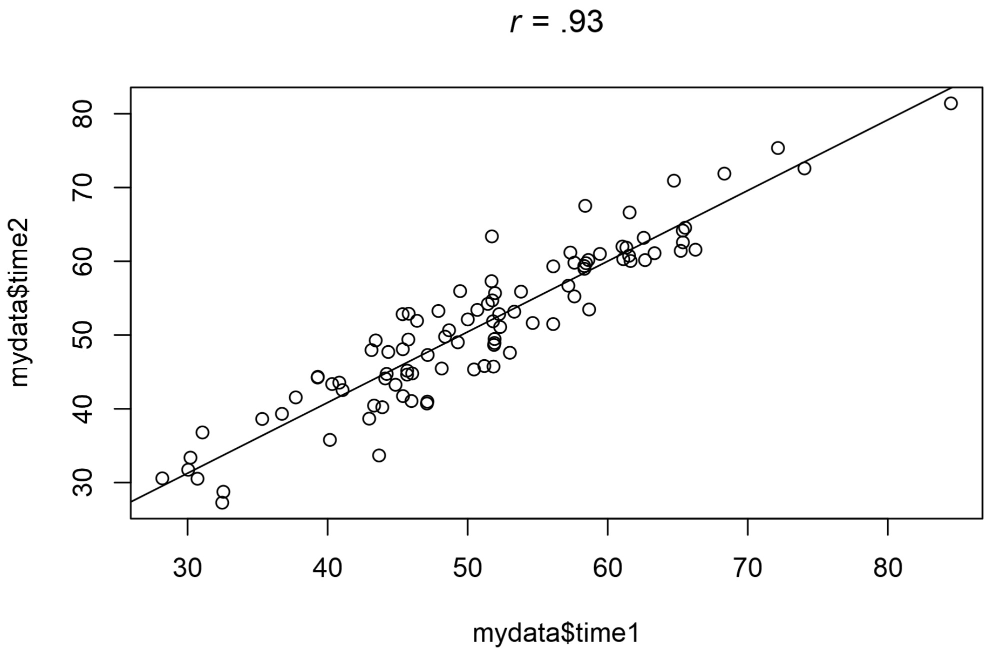


FIGURE 3.15 Test–Retest Reliability Scatterplot. The black line is the best-fitting linear line.

3.5.1.1.1 Considerations About the Correlation Coefficient

The correlation coefficient ranges from -1.0 to +1.0. The correlation coefficient (r) tells you two things: (1) the direction of the association (positive or negative) and (2) the magnitude of the association. If the correlation coefficient is positive, the association is positive. If the correlation coefficient is negative, the association is negative. If the association is positive, as X increases, Y increases (or conversely, as X decreases, Y decreases). If the association is negative, as X increases, Y decreases (or conversely, as X decreases, Y increases). The smaller the absolute value of the correlation coefficient (i.e., the closer the r value is to zero), the weaker the association and the flatter the slope of the best-fit line in a scatterplot. The larger the absolute value of the correlation coefficient (i.e., the closer the absolute value of the r value is to one), the stronger the association and the steeper the slope of the best-fit line in a scatterplot. See Figure 3.16 for a range of different correlation coefficients and what some example data may look like for each direction and strength of association.

Keep in mind that the Pearson correlation examines the strength of the *linear* association between two variables. If the association between two variables is nonlinear, the Pearson correlation provides the strength of the linear trend and may not provide a meaningful index of the strength of the association between the variables. For instance, Anscombe's quartet includes four sets of data that have nearly identical basic descriptive statistics (see Tables 3.1 and 3.2), including the same bivariate correlation, yet have very different distributions and whose association takes very different forms (see Figure 3.17).

TABLE 3.1 Anscombe's Quartet.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

TABLE 3.2 Descriptive Statistics of Anscombe's Quartet.

Property	Value
Sample size	11
Mean of X	9.0
Mean of Y	~7.5
Variance of X	11.0
Variance of Y	~4.1
Equation of regression line	Y = 3 + 0.5X
Standard error of slope	0.118
One-sample t-statistic	4.24

Property	Value
Sum of squares of X	110.0
Regression sum of squares	27.50
Residual sum of squares of Y	13.75
Correlation coefficient	.816
Coefficient of determination	.67

As an index of correlation, the coefficient of stability (and dependability) has important weaknesses. It is a form of *relative* reliability rather than *absolute* reliability: It examines the consistency of scores across time relative to variation across people. Higher stability coefficients reflect greater stability of individual differences—not greater stability in people's absolute level on the measure. This is a major limitation. Figure 3.18 depicts two example data sets that show strong relative reliability (a strong coefficient of stability) but poor absolute reliability based on inconsistency in people's absolute level across time.

Another limitation of the coefficient of stability (and dependability) is that it is sensitive to outliers. Additionally, if there are little or no individual differences in scores on a given measure, the coefficient of stability will be low relative to the true reliability because correlation coefficients tend to be attenuated in the presence of a restricted range, and it may not be a useful index of reliability depending on the purpose. For more information see Section 23.3.4.3.2 on the reliability paradox in the chapter on cognitive assessment.

See Figure 3.19 for an example of how a correlation coefficient tends to be attenuated when the range of one or both of the variables has restriction of range. The figure depicts the correlation with and without restriction of range on x (i.e., x is restricted to values between 55 and 65).

The observed correlation became much weaker due to restriction of range. Thus, when developing measures, it is important to ensure there is adequate variability of scores (i.e., individual differences) and that scores are not truncated due to ceiling or floor effects. Moreover, when interpreting associations with truncated variability, it is important to keep in mind that the true association is likely to be even stronger than what was observed if the measures did not have restricted range.

To address these limitations of the coefficient of stability, it is important to consider additional indices of test–retest reliability, such as the coefficient of repeatability and bias that are depicted in Bland-Altman plots, as described later.

3.5.1.2 Coefficient of Agreement

The intraclass correlation (ICC) can be used to evaluate the extent of absolute agreement of scores within a person across time. ICC ranges from 0–1, with higher scores indicating greater agreement. ICC was estimated using the `psych` package (Revelle, 2022).

```
ICC(mydata[,c("time1","time2")], missing = FALSE)
```

```
Call: ICC(x = mydata[, c("time1", "time2")], missing = FALSE)

Intraclass correlation coefficients
                         type  ICC   F df1 df2       p lower bound
Single_raters_absolute   ICC1 0.93 28  99 100 1.1e-45        0.90
```

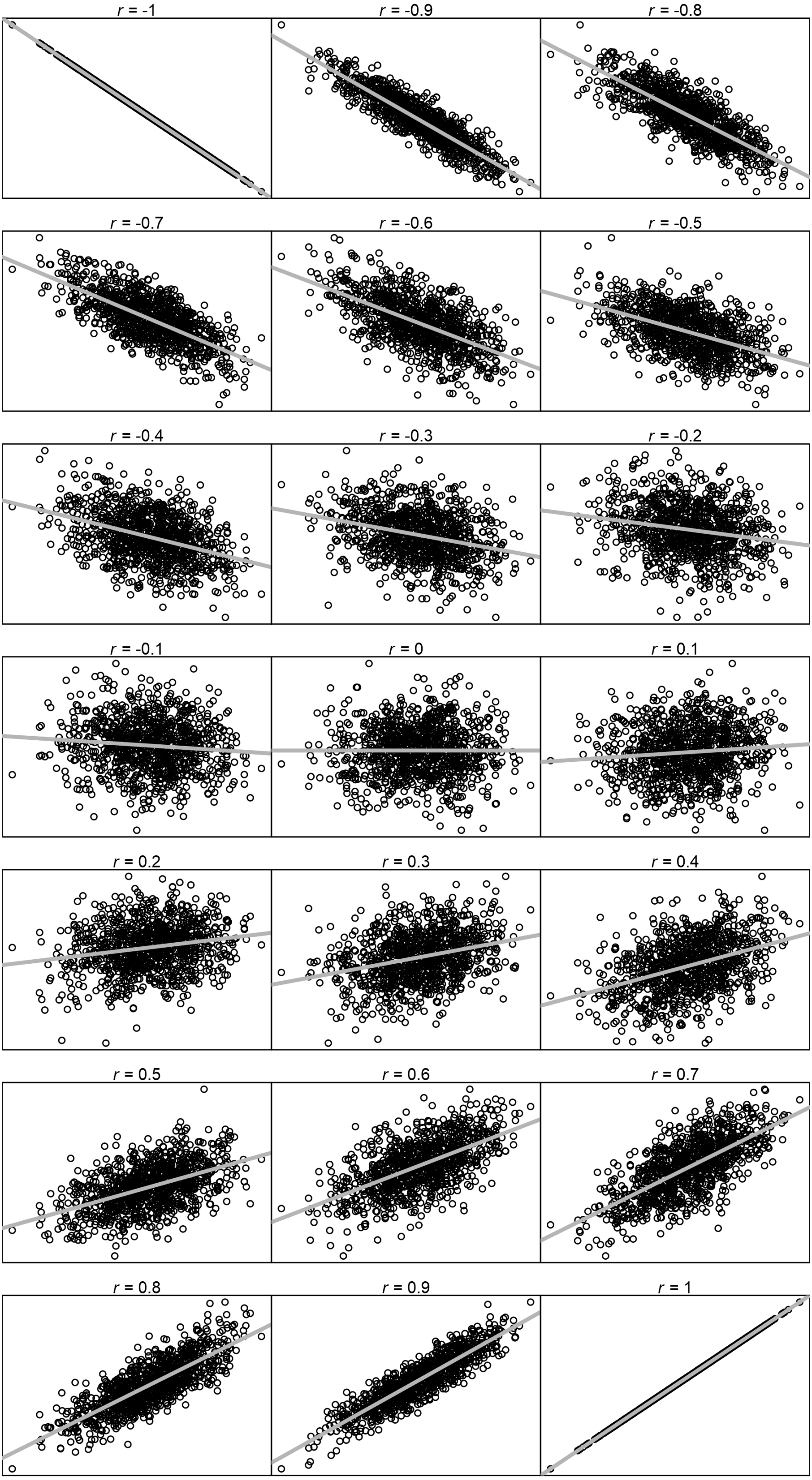


FIGURE 3.16 Correlation Coefficients.

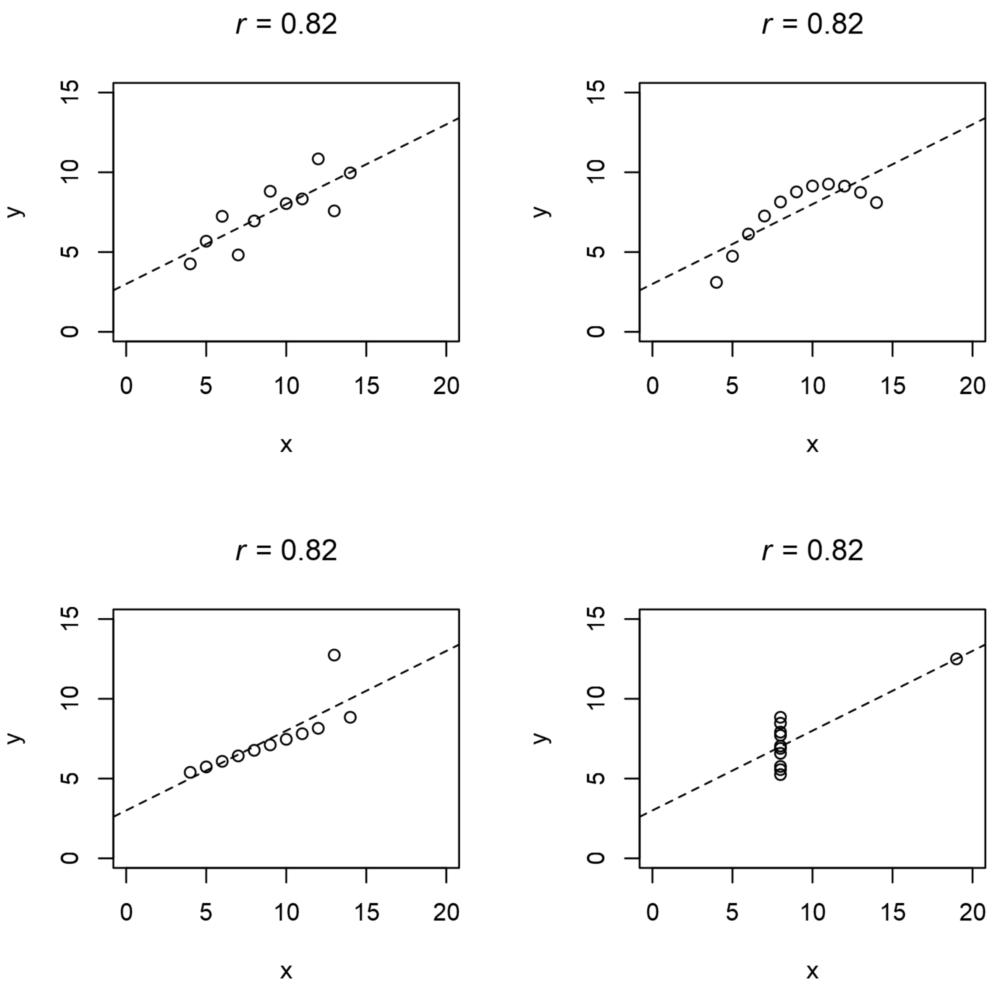


FIGURE 3.17 Anscombe's Quartet.

```
Single_random_raters     ICC2 0.93 28  99  99 2.6e-45        0.90
Single_fixed_raters      ICC3 0.93 28  99  99 2.6e-45        0.90
Average_raters_absolute ICC1k 0.96 28  99 100 1.1e-45        0.95
Average_random_raters   ICC2k 0.96 28  99  99 2.6e-45        0.95
Average_fixed_raters    ICC3k 0.96 28  99  99 2.6e-45        0.95
                        upper bound
Single_raters_absolute         0.95
Single_random_raters           0.95
Single_fixed_raters            0.95
Average_raters_absolute        0.98
Average_random_raters          0.98
Average_fixed_raters           0.98

 Number of subjects = 100     Number of Judges =  2
```

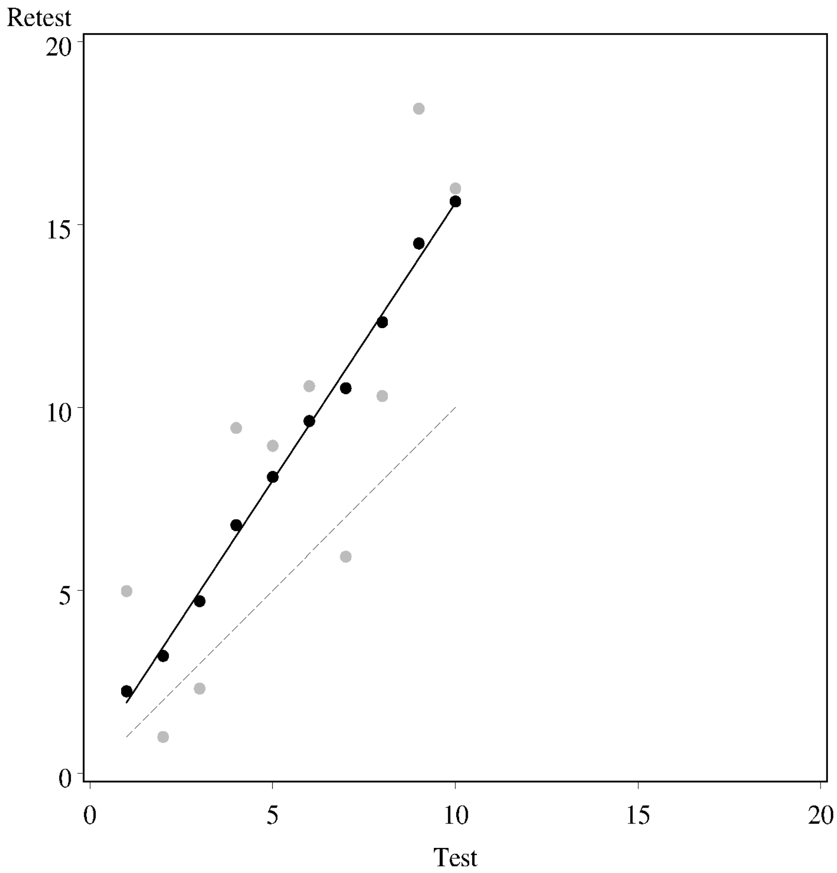


FIGURE 3.18 Hypothetical Data Demonstrating Good Relative Reliability Despite Poor Absolute Reliability. The figure depicts two fictional data sets (black and gray circles), which both exhibit a similar linear association. The line of best fit is the solid line in the graph and is the same for both data sets, but the black circles sit much closer to the line than the gray circles, leading to a much higher coefficient of stability (r = .99 and .84, respectively). However, neither sets of circles are on the line of complete agreement represented by the dashed line in the graph. If the circles fall on the line of complete agreement, it indicates that the measure's scores show consistency in absolute scores across time. Thus, although the measures show strong relative reliability, they show poor absolute reliability. (Figure reprinted from Vaz et al. (2013), figure 1, p. 3. Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS ONE*, *8*(9), e73990. https://doi.org/10.1371/journal.pone.0073990)

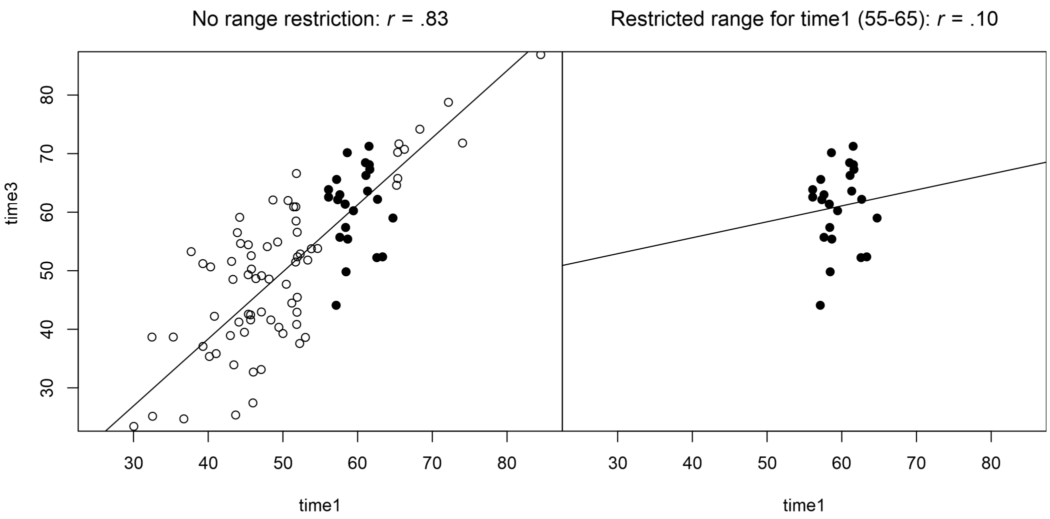


FIGURE 3.19 Example of Correlation With (Right Panel) and Without (Left Panel) Range Restriction. Filled black points represent the points in common across the two scatterplots.

```
See the help file for a discussion of the other 4 McGraw and Wong estimates,
```

There are various kinds of ICC coefficients. To read more about the various types of ICC coefficients, type the following command:

```
?ICC
```

3.5.1.3 Coefficient of Repeatability

The coefficient of repeatability (CR), also known as the "smallest real difference," is a form of *absolute* reliability. That is, it reflects the consistency of scores across time within a person. It is calculated as ~1.96 times the standard deviation of the mean differences between the two measurements (d_1 and d_2), as in Equation (3.14): https://www.medcalc.org/manual/bland-altman-plot.php (archived at https://perma.cc/MEH3-3TGN) (Bland & Altman, 1986, 1999). The CR is used for determining the lower and upper limits of agreement because it defines the range within which 95% of differences between the two measurement methods are expected to be.

The minimum possible coefficient of repeatability is zero, but there is no theoretical limit on the maximum possible coefficient of repeatability. The smaller the coefficient of repeatability (i.e., the closer to zero), the more consistent the scores are across time within a person. The coefficient of repeatability is tied to the units of the measure, so how small is considered "good" absolute reliability depends on the measure. In general, however, given the same units of measurement, a smaller coefficient of repeatability indicates stronger absolute test–retest reliability.

$$CR = 1.96 \times \sqrt{\frac{\sum(d_1 - d_2)^2}{n}} \tag{3.14}$$

The `petersenlab` package (Petersen, 2024) contains the `repeatability()` function that estimates the repeatability of a measure's scores and generates a Bland-Altman plot in Figure 3.20. The coefficient of repeatability is labeled `cr` in the output.

```
repeatability(
  measure1 = mydata$time1,
  measure2 = mydata$time2)
```

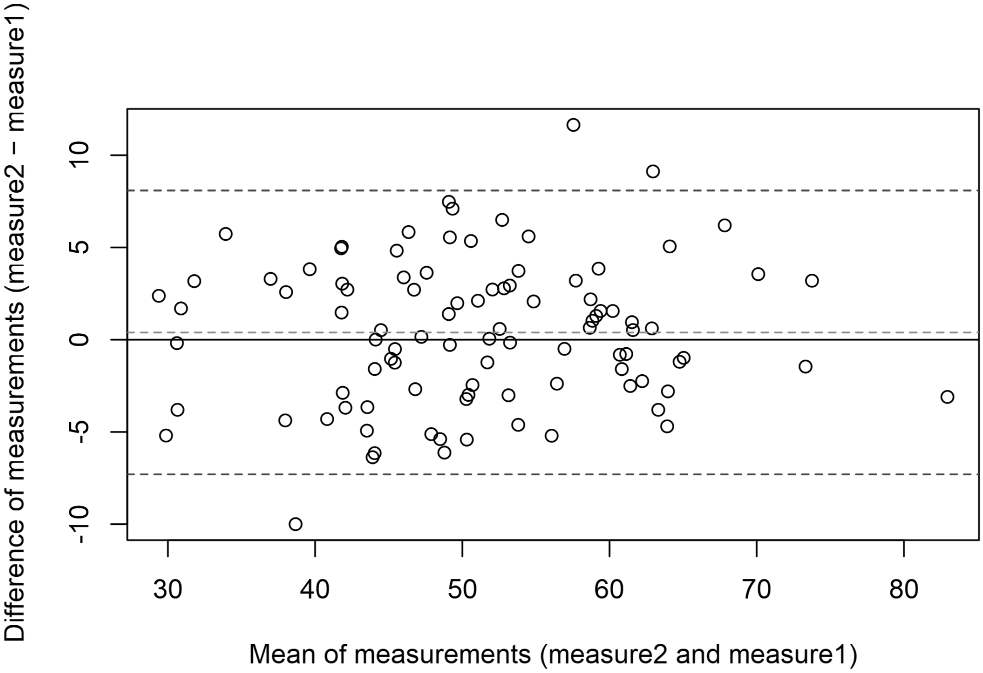


FIGURE 3.20 Bland-Altman Plot.

```
     cr   bias lowerLOA upperLOA
1 7.694 0.3956   -7.298     8.09
```

3.5.1.3.1 Bland-Altman Plot

The Bland-Altman plot can be useful for visualizing the degree of (in)consistency of people's absolute scores on a measure across two time points (Bland & Altman, 1986, 1999).

In a Bland-Altman plot, the x-axis represents the mean of a person's score across the two time points, and the y-axis represents the within-subject differences across time points. The horizontal line at $y = 0$ is called the line of identity. Points would fall on the line of identity if people's scores showed perfect consistency in their absolute level across time points (i.e., not relative level as in a cross-time correlation). Another line reflects the bias estimate (i.e., mean difference from the measure across the two points). The bias plus or minus the coefficient of repeatability is the upper or lower limit of agreement (LOA), respectively, as in Equations (3.15) and (3.16). The top dashed line reflects the upper LOA. The bottom dashed line reflects the lower LOA.

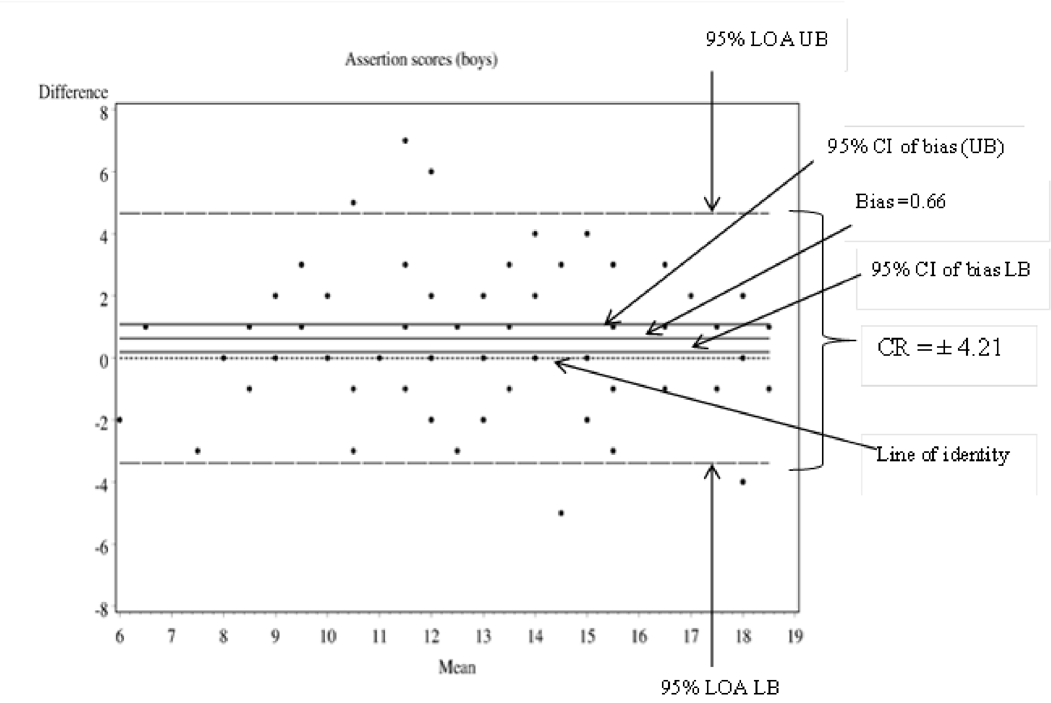


FIGURE 3.21 Example of a Bland-Altman Plot. (Figure reprinted from Vaz et al. (2013), figure 2, p. 4. Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS ONE*, *8*(9), e73990. https://doi.org/10.1371/journal.pone.0073990)

3.5.1.3.2 Bias

Bias is the degree of systematic measurement error. Bias is also called "mean error" in the context of assessing accuracy against a criterion (see Section 9.4.1.1.1). In a Bland-Altman plot, (test–retest) bias is estimated as the average of the within-subject differences between the two measurements. Estimated (test–retest) bias values that are closer to zero reflect stronger absolute test–retest reliability at the group-level. However, (estimated) bias values can still be close to zero while individuals show large differences because large positive and negative differences could cancel each other out. As a result, it is also important to examine other indices of test–retest reliability, including the coefficient of stability and the coefficient of repeatability.

The bias of 0.40 indicates that the scores at time 2 are somewhat larger, on average, than the scores at time 1.

3.5.1.3.3 95% Limits of Agreement

$$\text{Lower Limit of Agreement} = \text{Bias} - \text{Coefficient of Repeatability} \tag{3.15}$$

$$\text{Upper Limit of Agreement} = \text{Bias} + \text{Coefficient of Repeatability} \tag{3.16}$$

It is expected that the 95% limits of agreement (LOA) include 95% of differences between the two measurement methods (assuming the scores are normally distributed), based on how the coefficient of repeatability is calculated.

The limits of agreement indicate that 95% of scores at time 2 are expected to fall within $[-7.30, 8.09]$ of scores at time 1.

3.5.2 Inter-Rater Reliability

Inter-rater reliability is the consistency of scores across raters. For instance, two different raters may make a diagnostic decision for the same client, two different people may provide observational ratings of the same participant, or two different researchers may process psychophysiological data (e.g., electroencephalography, electrocardiography) from the same person. Consistency of scores can also be assessed within the same rater, which is known as *intra*-rater reliability and is described in Section 3.5.3. For assessing inter-rater reliability, intraclass correlation (ICC) is recommended for continuous data, such as ratings of a participant's mood. By contrast, Cohen's kappa (κ) is recommended for categorical variables, such as making a diagnostic decision whether to give someone a diagnosis of depression or not.

3.5.2.1 Continuous Data

Use intraclass correlation (ICC) for estimating inter-rater reliability of continuous data. ICC ranges from 0–1, with higher scores indicating greater agreement. ICC was estimated using the `psych` package (Revelle, 2022).

```
ICC(mydata[,c(
  "rater1continuous","rater2continuous","rater3continuous")],
  missing = FALSE)
```

```
Call: ICC(x = mydata[, c("rater1continuous", "rater2continuous", "rater3continuous")], 
    missing = FALSE)

Intraclass correlation coefficients 
                         type  ICC  F df1 df2       p lower bound
Single_raters_absolute   ICC1 0.81 14  99 200 2.7e-54        0.75
Single_random_raters     ICC2 0.81 14  99 198 6.4e-54        0.75
Single_fixed_raters      ICC3 0.81 14  99 198 6.4e-54        0.75
Average_raters_absolute ICC1k 0.93 14  99 200 2.7e-54        0.90
Average_random_raters   ICC2k 0.93 14  99 198 6.4e-54        0.90
Average_fixed_raters    ICC3k 0.93 14  99 198 6.4e-54        0.90
                        upper bound
Single_raters_absolute         0.86
Single_random_raters           0.86
Single_fixed_raters            0.86
Average_raters_absolute        0.95
Average_random_raters          0.95
Average_fixed_raters           0.95

 Number of subjects = 100     Number of Judges =  3
See the help file for a discussion of the other 4 McGraw and Wong estimates,
```

3.5.2.2 Categorical Data

For estimating the inter-rater reliability of categorical data, use Cohen's kappa (Bakeman & Goodman, 2020), Fleiss's kappa, the S index (Falotico & Quatto, 2010), or Gwet's AC1 statistic (Gwet, 2008; Gwet, 2021a, 2021b). Cohen's kappa was estimated using the `psych` package (Revelle, 2022). Fleiss's kappa and Gwet's AC1 statistic were estimated using the `irrCAC` package. These estimates of inter-rater reliability range from −1 to +1, with −1 indicating perfect disagreement, 0 indicating chance agreement, and 1 indicating perfect agreement. Larger, more positive scores indicate greater agreement.

```
cohen.kappa(mydata[,c(
  "rater1categorical","rater2categorical","rater3categorical")])
```

```
Cohen Kappa (below the diagonal) and Weighted Kappa (above the diagonal)
For confidence intervals and detail print with all=TRUE
                  rater1categorical rater2categorical rater3categorical
rater1categorical              1.00              0.86              0.90
rater2categorical              0.86              1.00              0.76
rater3categorical              0.90              0.76              1.00

Average Cohen kappa for all raters  0.84
Average weighted kappa for all raters  0.84
```

```
fleiss.kappa.raw(na.omit(mydata[,c(
  "rater1categorical","rater2categorical","rater3categorical")]))
```

```
$est
     coeff.name     pa     pe coeff.val coeff.se      conf.int p.value
1 Fleiss' Kappa 0.9175 0.5001     0.835  0.04483 (0.746,0.924)       0
      w.name
1 unweighted

$weights
     [,1] [,2]
[1,]    1    0
[2,]    0    1

$categories
[1] 0 1
```

```
gwet.ac1.raw(na.omit(mydata[,c(
  "rater1categorical","rater2categorical","rater3categorical")]))
```

```
$est
  coeff.name     pa     pe coeff.val coeff.se      conf.int p.value
1        AC1 0.9175 0.4999    0.8351  0.04479 (0.746,0.924)       0
      w.name
1 unweighted

$weights
```

```
     [,1] [,2]
[1,]    1    0
[2,]    0    1

$categories
[1] 0 1
```

3.5.3 Intra-Rater Reliability

Principles of intra-rater reliability are similar to principles of inter-rater reliability (described in Section 3.5.2), except the *same* rater provides ratings on two occasions, rather than using multiple raters.

3.5.4 Parallel- (or Alternate-) Forms Reliability

Parallel-forms reliability (also known as alternate-forms reliability) is the consistency of scores across two parallel forms. Parallel forms are two equivalent measures of the same construct that differ in content or format. The two measures can be administered in the same occasion, known as *immediate parallel-forms reliability*, or they can be administered separated by a delay, known as *delayed parallel-forms reliability*. Similar considerations of length of delay duration and ways to reduce carryover effects are relevant to parallel-forms reliability as they are to test–retest reliability.

Parallel-forms reliability is often estimated with the Pearson correlation, which is known as the *coefficient of equivalence*. The coefficient of equivalence is the consistency of scores across two parallel forms relative to variation across people. It is interpreted as the ratio of true score variance to observed score variance, consistent with the formal definition of reliability.

An example of parallel-forms reliability would be if we want to assess a participant twice with different measures of the same construct to avoid practice effects. This approach is often conducted in academic achievement testing, and it could involve retesting that is immediate or delayed. For instance, we could administer item set A at time 1 and item set B at time 2 for the same participant, to get rid of any possibility that the person's score differed across time because of improved performance that resulted merely from prior exposure to the same items. Many standardized academic achievement and aptitude tests have developed parallel forms, including the SAT, ACT, and GRE.

Parallel-forms reliability controls for specific error, i.e., error that is particular to a specific measure. However, parallel-forms reliability has the limitation that it assumes that the parallel forms are equivalent, such that it makes no difference which test you use. Two forms are considered equivalent when they assess the same construct, have the same mean and variance of scores, and have the same inter-correlations with external criteria. Technically, two instruments are only truly equivalent if they have the same true scores and variability of error scores, though this is difficult to establish empirically. Given these constraints, it can be difficult to assume that different forms are equivalent. Nevertheless, integrative data analysis and measurement harmonization may help make the assumption more tenable (Hussong et al., 2013, 2020). In general, parallel forms can be difficult to create and they can more than double the time it takes to develop a single measure, often with limited benefit, depending on the intended use.

3.5.5 Internal Consistency Reliability

Internal consistency reliability is the consistency of scores across items, that is, their interrelatedness. Internal consistency is necessary, but insufficient, for establishing the homogeneity or unidimensionality of the measure, i.e., that the items on the measure assess the same construct.

There are many examples of internal consistency estimates, including Cronbach's coefficient alpha (α), Kuder-Richardson Formula 20 (which is a special case of Cronbach's alpha with dichotomous items), omega coefficients, average variance extracted, greatest lower bound, coefficient H, split-half reliability, average inter-item correlation, and average item–total correlation. In general, internal consistency ranges from 0–1, with higher scores indicating greater consistency. There is no magic cutoff for determining adequate internal consistency; however, values below .70 suggest that the items may not assess the same construct.

Internal consistency can be affected by many factors, such as

- wording of the item
- a participant's comprehension of the item, which could be impacted if using items with dated language or double negatives
- interpretation
- attention of the participant
- compliance of the participant

Although establishing internal consistency can help ensure that items in the measure are inter-related, you do not want to select items just based on internal consistency. If you did, you would end up with items that are too similar. For example, you could end up only with items that have similar item stems and are coded in the same direction. Therefore, selecting items solely based on internal consistency would likely not assess the breadth of the construct and would reflect strong method variance of the wording of the particular items. In other words, internal consistency can be *too* high. Internal consistency is helpful to a point, but you do not want items to be too homogeneous or redundant.

3.5.5.1 Average Inter-Item Correlation

One estimate of internal consistency is the average (mean or median) inter-item correlation.

```
interItemCorrelations <- cor(
  mydata[,c("item1","item2","item3","item4")],
  use = "pairwise.complete.obs")
diag(interItemCorrelations) <- NA
```

Here is the mean inter-item correlation for each item:

```
meanInterItemCorrelations <- colMeans(
  interItemCorrelations, na.rm = TRUE)
meanInterItemCorrelations
```

```
 item1  item2  item3  item4 
0.7748 0.7787 0.7539 0.6595 
```

Here is the mean inter-item correlation across all items:

```
psych::alpha(
  mydata[,c("item1","item2","item3","item4")])$total$average_r
```

```
[1] 0.7417
```

Here is the median inter-item correlation for each item:

```
medianInterItemCorrelations <- apply(
  interItemCorrelations, 2,
  function(x) median(x, na.rm = TRUE))
medianInterItemCorrelations
```

```
 item1  item2  item3  item4 
0.7578 0.7792 0.7578 0.6317 
```

Here is the median inter-item correlation across all items:

```
psych::alpha(mydata[,c(
  "item1","item2","item3","item4")])$total$median_r
```

```
[1] 0.7412
```

3.5.5.2 Average Item–Total Correlation

Another estimate of internal consistency is the average item–total correlation.

```
itemTotalCorrelations <- psych::alpha(
  mydata[,c("item1","item2","item3","item4")])$item.stats["raw.r"]

mean(itemTotalCorrelations$raw.r)
```

```
[1] 0.894
```

When examining the item–total correlation for a given item, it is preferable to exclude the item of interest from the total score, so that the item is associated with the total score that is calculated from the remaining items. This reduces the extent to which the association reflects that the item is associated with itself.

```
itemTotalCorrelationsDrop <- psych::alpha(
  mydata[,c("item1","item2","item3","item4")])$item.stats["r.drop"]

mean(itemTotalCorrelationsDrop$r.drop)
```

```
[1] 0.8053
```

3.5.5.3 Average Variance Extracted

The equivalent of the average item–total (squared) correlation when dealing with a reflective latent construct is the average variance extracted (AVE). AVE is estimated as the mean of the squared loadings of the indicators of a latent factor (i.e., the sum of the squared loadings divided by the number of indicators). When using unstandardized estimates, AVE is the item-variance weighted average of item reliabilities (Rönkkö & Cho, 2020). When using standardized estimates, AVE is the average indicator reliability (Rönkkö & Cho, 2020). AVE is equivalent to the communality of a latent factor. AVE was estimated using a confirmatory factor analysis model using the `lavaan` (Rosseel et al., 2022) and `semTools` (Jorgensen et al., 2021) packages.

```
latentFactorModel_syntax <- '
 latentFactor =~ item1 + item2 + item3 + item4
'

latentFactorModel_fit <- cfa(
  latentFactorModel_syntax,
  data = mydata,
  missing = "ML",
  estimator = "MLR",
  std.lv = TRUE)
```

```
AVE(latentFactorModel_fit)
```

```
latentFactor
        0.64
```

3.5.5.4 Split-Half Reliability

When estimating split-half reliability, we randomly take half of the items on the measure and relate scores on these items to the score on the other half of items. One challenge to split-half reliability is that, according to classical test theory, shorter tests tend to be less reliable than longer tests, so this would be an underestimate of the test's reliability. In general, the greater the number of items, the more reliable the test. To overcome this, the Spearman-Brown prediction (or prophecy) formula predicts the reliability of a test after changing the test length. Thus, the Spearman-Brown prediction formula accounts for the shorter test in predicting what the reliability of the full-length test would be when estimating split-half reliability.

The Spearman-Brown prediction formula is in Equation (3.17):

$$\text{predicted test reliability, } r_{xx} = \frac{n \cdot r_{xx'}}{1 + (n-1)r_{xx'}} \tag{3.17}$$

where n = the number of tests combined (i.e., the ratio of the number of new items to the number of old items) and $r_{xx'}$ = the reliability of the original test.

So, according to Equation (3.18), doubling the length of a test whose reliability is .80 would be estimated to have a reliability of

$$\begin{aligned} \text{predicted test reliability, } r_{xx} &= \frac{2 \cdot .80}{1 + (2-1).80} \\ &= .89 \end{aligned} \tag{3.18}$$

Another challenge to split-half reliability is that the reliability differs depending on which items go into which half.

One option is to calculate the split-half reliability of odd and even trials, while correcting for test length, as is performed with the `item_split_half()` function of the `performance` package (Lüdecke et al., 2021):

```
item_split_half(mydata[,c(
  "item1","item2","item3","item4")])
```

```
$splithalf
[1] 0.8734

$spearmanbrown
[1] 0.9324
```

Modern computational software allows for estimating the mean reliability estimate of all possible split-halves (Revelle & Condon, 2019), while correcting for test length, as seen below. Split-half reliability was estimated using the `psych` package (Revelle, 2022).

```
splitHalf(mydata[,c(
  "item1","item2","item3","item4")],
  brute = TRUE)
```

```
Split half reliabilities
Call: splitHalf(r = mydata[, c("item1", "item2", "item3", "item4")],
    brute = TRUE)

Maximum split half reliability (lambda 4) =  0.95
Guttman lambda 6                          =  0.93
Average split half reliability            =  0.92
Guttman lambda 3 (alpha)                  =  0.92
Guttman lambda 2                          =  0.92
Minimum split half reliability  (beta)    =  0.87
Average interitem r =  0.75  with median =  0.74
```

3.5.5.5 Cronbach's Alpha (Coefficient Alpha)

Cronbach's alpha (α), also known as coefficient alpha, is approximately equal to the mean reliability estimate of all split-halves, assuming that item standard deviations are equal. Alpha considers the inter-relatedness of a total set of items—giving us information about the extent to which each item in a set of items correlates with all other items. Alpha is the most commonly used estimate for internal consistency reliability, but it has many problems.

Cronbach's alpha assumes that the items are

- unidimensional: that there is one predominant dimension that explains the variance among the items
- essentially tau (τ) equivalent: that the items are equally related to the construct

Cronbach's alpha is affected by the

- number of items: The more items, the more inflated alpha is, even if correlations among the items are low.
- variance of the item scores: If the variances of item scores are low (e.g., dichotomous scores), alphas tend to be an underestimate of internal consistency reliability.

- violations of assumptions:
  - Cronbach's alpha assumes items are essentially tau equivalent—that the items are equally related to the construct. However, this is an unrealistic assumption.
  - Cronbach's alpha depends on the dimensionality of the scale: Alpha is only valid when the scale is unidimensional. If you use it with multidimensional scales, you know the extent to which items correlate with another but nothing about whether they are actually measuring the same constructs.

Cronbach's alpha is appropriate only in situations where a scale is unidimensional, where the items equally contribute to the scale, and you would like to determine whether its items are inter-related. However, Cronbach's alpha does not assess the dimensionality of the scale or any type of construct validity—it can only tell us how strongly the items hang together.

Cronbach's alpha (α) was estimated using the `psych` package (Revelle, 2022).

```
psych::alpha(mydata[,c(
  "item1","item2","item3","item4")])$total$raw_alpha
```

```
[1] 0.8924
```

However, because of the many weaknesses of Cronbach's alpha (Cortina, 1993; Dunn et al., 2014; Flora, 2020; Green & Yang, 2015; A. F. Hayes & Coutts, 2020; Kelley & Pornprasertmanit, 2016; McNeish, 2018; Peters, 2014; Raykov, 2001; Revelle & Condon, 2019; Sijtsma, 2008; but see Raykov & Marcoulides, 2019), current recommendations are to use coefficient omega (ω) instead, as described next.

3.5.5.6 Coefficient Omega (ω)

Current recommendations are to use McDonald's coefficient omega (ω) instead of Cronbach's alpha (α) for internal consistency reliability. Specifically, it is recommended to use omega total (ω_t) for continuous items that are unidimensional, omega hierarchical (ω_h) for continuous items that are multidimensional, and omega categorical (ω_C) for categorical items (i.e., items with fewer than five ordinal categories), with confidence intervals calculated from a bias-corrected bootstrap (Flora, 2020; Kelley & Pornprasertmanit, 2016).

3.5.5.6.1 Omega total (ω_t)

Omega total (ω_t) is the proportion of the total variance that is accounted for by the common factor (Flora, 2020). Omega total, like Cronbach's alpha, assumes that the items are unidimensional. Omega total (ω_t) was calculated using the `MBESS` package (Kelley, 2020). This code calculates a point estimate for omega total:

```
ci.reliability(
  mydata[,c("item1","item2","item3","item4")],
  type = "omega",
  interval.type = "none")$est
```

```
[1] 0.8761
```

3.5.5.6.2 Omega hierarchical (ω_h)

If the items are multidimensional and follow a bifactor structure, omega hierarchical (ω_h) is preferred over omega total. Omega hierarchical is the proportion of the total variance that is accounted for by the general factor in a set of multidimensional items (Flora, 2020). Use omega hierarchical for estimating internal consistency of continuous items that are multidimensional and that follow a bifactor structure. If the multidimensional items follow a higher-order structure, use omega higher-order [ω_{ho}; Flora (2020)].

Omega hierarchical was calculated using the `MBESS` package (Kelley, 2020).

```
ci.reliability(
  mydata[,c("item1","item2","item3","item4")],
  type = "hierarchical",
  interval.type = "none")$est
```

```
[1] 0.8453
```

3.5.5.6.3 Omega categorical (ω_C)

Use omega categorical (ω_C) for categorical items that are unidimensional. To handle categorical data, omega categorical is calculated using a polychoric correlation matrix instead of a Pearson correlation matrix. If the categorical items are multidimensional items and they follow a bifactor structure, use omega hierarchical categorical [$\omega_{h\text{-cat}}$; Flora (2020)]. If the categorical items are multidimensional and they follow a higher-order structure, use omega higher-order categorical [$\omega_{ho\text{-cat}}$; Flora (2020)].

Omega categorical was calculated using the `MBESS` package (Kelley, 2020).

```
ci.reliability(
  mydata[,c(
    "rater1categorical",
    "rater2categorical",
    "rater3categorical")],
  type = "categorical",
  interval.type = "none")$est
```

```
       [,1]   [,2]   [,3]
[1,] 1.0073 0.9852 0.9904
[2,] 0.9852 0.9636 0.9686
[3,] 0.9904 0.9686 0.9737
```

```
[1] 0.9601
```

3.5.6 Comparison

Even though the types of reliability described in Section 3.5 are all called "reliability," they are very different from each other. Test–retest reliability and inter-rater reliability tend to be lower than immediate parallel-forms reliability and internal consistency reliability because they involve administering measures at different times or with different raters. However, delayed parallel-forms reliability tends to be lower than test–retest reliability because it involves assessing the construct at a different time point *and* with different items. Thus, if we put the different types of reliability in order from low to high, they tend to be (but not always) arranged like this (Equation (3.19)):

$$\begin{aligned}
\text{delayed parallel-forms} &< \\
\text{test–retest, inter-rater} &< \\
\text{immediate parallel-forms, internal consistency, intra-rater}&
\end{aligned} \tag{3.19}$$

3.5.7 Reliability of a Linear Composite Score

A linear composite score is a linear combination of variables, each of which has a given weight in the linear composite. Unit weights (i.e., weights of 1) would reflect an unweighted composite.

We first specify the variables used to generate the correlation matrix, and the reliability, standard deviation, and weight of each variable. We specify items 2 and 3 to have twice the weight as item 1 in the composite, and item 4 to have 1.5 times the weight as item 1 in the composite.

```r
rxx <- c(0.70, 0.75, 0.80, 0.85) #reliability coefficients
rho <- cor(mydata[,c("item1","item2","item3","item4")],
           use = "pairwise.complete.obs")
sigma <- apply(
  mydata[,c("item1","item2","item3","item4")],
  2,
  function(x) sd(x, na.rm = TRUE)) #standard deviations
weights <- c(1, 2, 2, 1.5) #weights of each variable in the linear combination
```

Here is the reliability of an unweighted composite:

```r
composite_rel_matrix(
  rel_vec = rxx,
  r_mat = rho,
  sd_vec = sigma)
```

```
[1] 0.9337
```

Here is the reliability of a weighted composite:

```r
composite_rel_matrix(
  rel_vec = rxx,
  r_mat = rho,
  sd_vec = sigma,
  wt_vec = weights)
```

```
[1] 0.9317
```

3.5.8 Reliability of a Difference Score

Difference scores tend to be lower in reliability than the individual indices that compose the difference score. This is especially the case when the two indices are correlated. If the correlation between the indices is equal to the reliability of the indices, the reliability of the difference score will be zero (Revelle & Condon, 2019). If the correlation between the two indices is large, it requires a very high reliability for the difference score to show a strong reliability (Trafimow, 2015). To the extent that the two indices have unequal variances (e.g., if one index has a standard deviation that is four times the standard deviation of the second index), the reliabilities do not need to be quite as high for the difference score to be reliable (Trafimow, 2015).

The reliability of a difference score can be estimated using Equation (3.20) (Trafimow, 2015):

$$r_{x-y,x-y} = \frac{\sigma^2_x r_{xx} + \sigma^2_y r_{yy} - 2r_{xy}\sigma_x\sigma_y}{\sigma^2_x + \sigma^2_y - 2r_{xy}\sigma_x\sigma_y} \tag{3.20}$$

The `petersenlab` package (Petersen, 2024) contains the `reliabilityOfDifferenceScore()` function to estimate the reliability of a difference score:

```
reliabilityOfDifferenceScore(
  x = mydata$item1,
  y = mydata$item2,
  reliabilityX = .95,
  reliabilityY = .95)
```

```
[1] 0.2553
```

For instance, in this case, the reliability of the difference score is .26, even though the reliability of both indices is .95.

In Figures 3.22 and 3.23, the reliability of a difference score is depicted as a function of (a) the reliability of the indices that compose the difference score and (b) the correlation between them, assuming that both indices have equal variances.

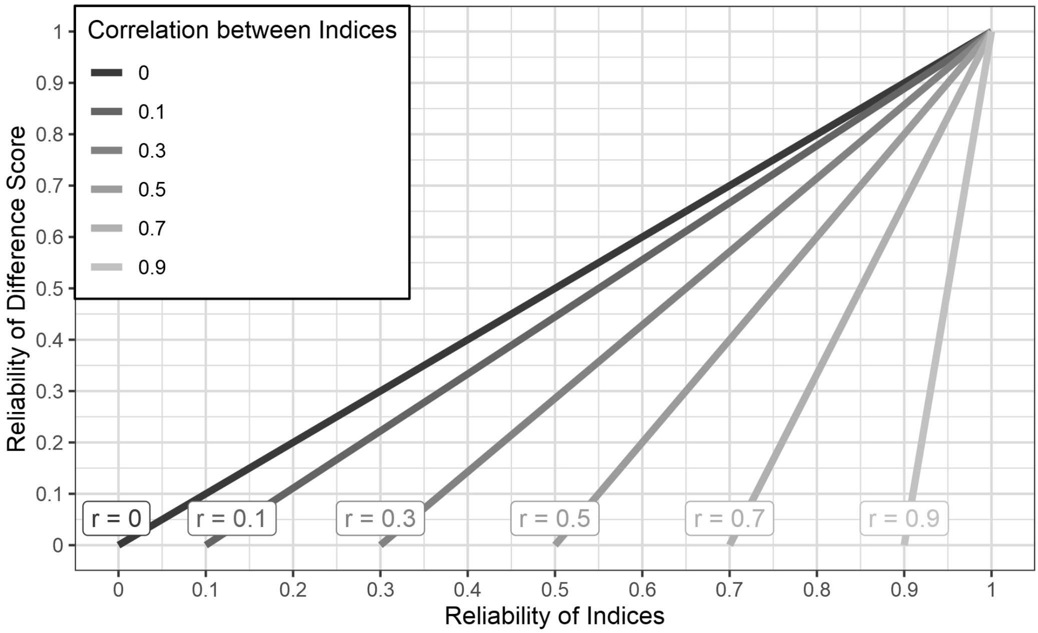


FIGURE 3.22 Reliability of Difference Score as a Function of Reliability of Indices and the Correlation Between Them.

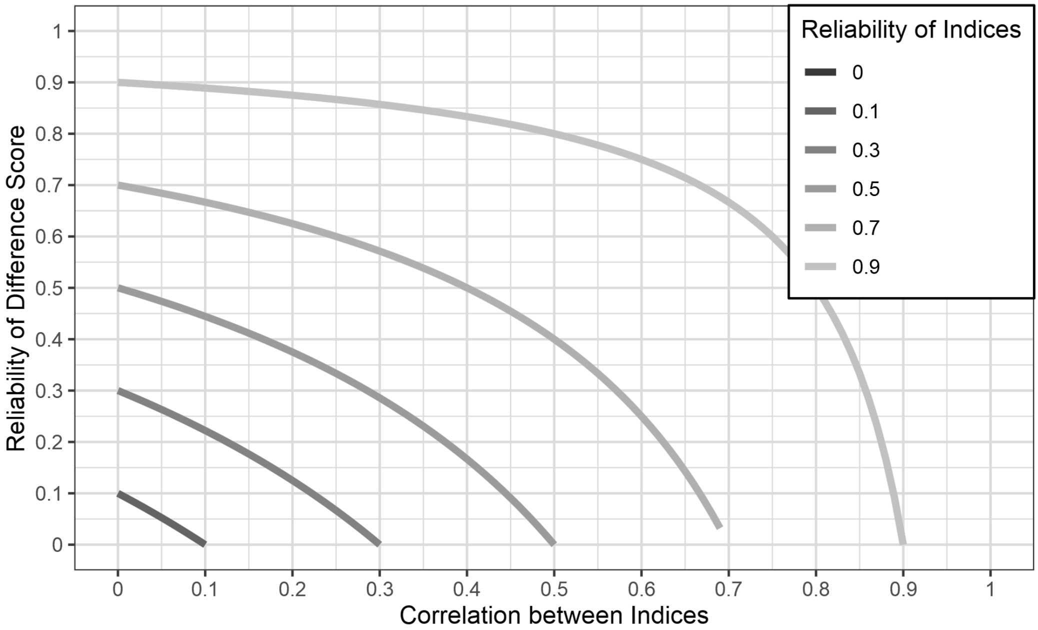


FIGURE 3.23 Reliability of Difference Score as a Function of Correlation Between Indices and Reliability of Indices.

3.6 Applied Examples

Although reliability is an important consideration for measures, stronger reliability is not always better. Reliability is not one thing; it is not monolithic. There are multiple aspects of reliability, and which aspects of reliability are most important depend on the construct of interest. That is, we want different kinds of reliability for measures of different constructs. In general, we want our measures to show strong inter-rater, intra-rater, immediate parallel-forms, and internal consistency reliability. However, we do not want internal consistency reliability to be too high so as to have redundant items that are correlated for reasons of shared method biases. Moreover, for some constructs (i.e., formative constructs such as stressful life events), we would not expect internal consistency reliability. And the extent to which our measures should show test–retest and delayed parallel-forms reliability depends on the stability of the construct and the time lag.

Consider the construct of reactive dysphoria. Assume that we assess people's reactive dysphoria with observations and several questionnaires today and one month from now. However, test–retest reliability is not meaningful because the phenomenon is not meant to be stable across time—it is supposed to be reactive to the situation, not a trait. Instead, we would expect that a strong measure of reactive dysphoria has high internal consistency reliability, a high coefficient of equivalence from immediate parallel-forms reliability, and high inter-/intra-rater reliability.

Consider the example of personality and intelligence. Personality is defined such that personality "traits" are thought to be stable across time and situation. Likewise, individual differences in intelligence are thought to be relatively stable in terms of individual differences across time (not in absolute level). Thus, strong measures of personality and intelligence should show high test–retest reliability, parallel-forms reliability, internal consistency reliability, and inter-/intra-rater reliability.

Just because one aspect of reliability is strong does not necessarily mean that other aspects of reliability will be strong. Measures of some constructs (e.g., depression) could be expected to show low test–retest reliability but high internal consistency reliability. You can also have measures that have low internal consistency reliability but that have high test–retest reliability (e.g., if our nonsensical measure averaged together z-scores of height and blood type).

Consider a questionnaire assessing people's extraversion. We would expect high stability of construct, and therefore a high coefficient of stability, but test and retest scores might not be equal. That is, the measures may show a somewhat lower repeatability coefficient because of random error—on a given day, the participant may be sick, stressed, or tired; they may have recently gotten into an argument; or they may be impacted by the season, the weather, or differences in the context (e.g., lab or online), just to name a few potential sources of random error.

3.7 Standard Error of Measurement

We presented and discussed the formula for estimating the standard error of measurement earlier in Equation (3.11).

The reliability of the measure was estimated here based on the test–retest reliability of the measure at T1 with the measure at T2.

```
reliabilityOfMeasure <- as.numeric(
  cor.test(
    x = mydata$time1,
    y = mydata$time2)$estimate)

sem <- sd(mydata$time1, na.rm = TRUE) * sqrt(1 - reliabilityOfMeasure)
sem
```

```
[1] 2.72
```

We can use the standard error of measurement to estimate the confidence interval around a person's score. A confidence interval indicates a range within which, with some degree of confidence (e.g., 95%), the person's true score is expected to be. For a 95% confidence interval, the true score is expected to fall within ±1.96 SEMs of the observed score 95% of the time.

```
lower95CI <- mean(mydata$time1, na.rm = TRUE) - qnorm(.975)*sem
upper95CI <- mean(mydata$time1, na.rm = TRUE) + qnorm(.975)*sem

lower95CI
```

```
[1] 45.6
```

```
upper95CI
```

```
[1] 56.27
```

The 95% CI is $[45.6, 56.27]$.

3.8 Influences of Measurement Error on Test–Retest Reliability

This section describes the influence of measurement error on reliability, specifically test–retest reliability. There are different types of measurement error, and each can have different impacts on reliability and, ultimately, your findings. As discussed in Section 3.2, a given error can be either systematic or random. In addition, error can be within-person, between-person, or both. However, error scores can include multiple errors simultaneously.

3.8.1 Add within-person random error

To add within-person random error, we added error that is randomly distributed around a mean of 0 and a standard deviation of 3 to all participants' scores at all time points. Adding within-person random error led to

- a weaker (lower) stability coefficient
- a weaker (higher) repeatability coefficient

3.8.2 Add within-person systematic error

To add within-person systematic error to all participants, we added error that is consistent across time for a given participant: pulled from a distribution with a mean of 5 and a standard deviation of 3. Adding within-person systematic error led to

- an artificially stronger (higher) stability coefficient
- no effect on the repeatability coefficient
- bias at the person level, but this bias would go undetected by the estimate of group-level bias in the Bland-Altman plot

3.8.3 Add between-person random error at T2

To add between-person random error, we added error that is randomly distributed around a mean of 0 and a standard deviation of 3 to all participants' scores at time 2 (see below). Adding between-person random error led to

- a weaker (lower) stability coefficient
- a weaker (higher) repeatability coefficient

3.8.4 Add between-person systematic error at T2

To add between-person systematic error, we added constant variance of 10 to all participants' scores at T2. Adding between-person systematic error led to

- no effect on the stability coefficient

- no effect on the repeatability coefficient
- bias at the person level and at the group level (detected by group-level estimate of bias in Bland-Altman plot)

3.8.5 Add constant variance of 10 to all participants at all time points

Adding constant variance (10) to all participants' scores at all time points led to

- no effect on the stability coefficient
- no effect on the repeatability coefficient
- bias at the person level and at the group level, but this bias would go undetected by the estimate of group-level bias in the Bland-Altman plot

3.8.6 Summary

To summarize the effects of measurement error, random error (both within- and between-person) results in (a) a worse (lower) stability coefficient because a smaller proportion of the observed score variance is attributable to true score variance, and (b) a worse (higher) repeatability coefficient because the standard deviation of the difference between the two measurements is greater. Neither type of random error leads to bias at the person or group level.

In terms of systematic error, all types of systematic error lead to increased bias, i.e., reduced accuracy for the person-level average and/or group-level average. Though multiple biases can coexist and can work in opposite directions, which can obscure bias even though it exists. In terms of within-person systematic error, it (a) artificially increases the stability coefficient, even though the measure is not actually more reliable, (b) has no effect on the repeatability coefficient because it is not changing the difference between the two measurements for a person, and (c) results in bias at the person level because the person's mean systematically differs from their true score.

In terms of between-person systematic error at time 2 (an example of adding constant variance to all participants), it (a) had no effect on the stability coefficient because the rank order of individual differences stays the same, (b) had no effect on the repeatability coefficient because the standard deviation of the difference between the two measurements is not changed, (c) resulted in bias at the person level and at the group level that would be detected in a Bland-Altman plot because the group-level average deviates from the group's average true score.

In sum, just because a measure has high estimates of test–retest reliability does not mean that the measure is actually reliable because systematic error does not reduce the stability and repeatability coefficients. Indeed, within-person systematic error actually (and artificially) *increases* the coefficient of stability. The coefficient of stability index of test–retest reliability is affected by the measure *and* variance across people, and it can be thrown off by outliers or by range restriction because the Pearson correlation strongly depends on these, as discussed in Section 3.5.1.1.1. And sample variation also has a strong influence, so it is important to cross-validate estimates of reliability in independent samples.

3.9 Effect of Measurement Error on Associations

As we describe in greater detail in the chapter on validity in Section 4.6, (random) measurement error weakens (or attenuates) the association between variables. The greater the random measurement error, the weaker the association.

3.9.1 Attenuation of True Test–Retest Reliability Due to Measurement Error

In the context of examining the association between test and retest (i.e., test–retest reliability), the estimate of the stability of the construct (i.e., true test–retest association) is attenuated to the extent that the measure is unreliable, as described in Equation (3.21):

$$\begin{aligned} r_{x_1 x_2} &= r_{x_{1_t} x_{2_t}} \sqrt{r_{xx} r_{xx}} \\ &= r_{x_{1_t} x_{2_t}} \sqrt{{r_{xx}}^2} \\ &= r_{x_{1_t} x_{2_t}} r_{xx} \\ \text{test–retest correlation of measure} &= (\text{true correlation of construct across two time points}) \\ &\quad \times (\text{reliability of measure}) \end{aligned} \tag{3.21}$$

Find the observed test–retest association if the true test–retest association (i.e., the stability of the construct) is .7 and the reliability of the measure is .9. The `petersenlab` package (Petersen, 2024) contains the `attenuationCorrelation()` function that estimates the observed association given the true association and the reliability of the predictor and criterion. When dealing with test–retest reliability, the predictor and criterion are the same, so we use the same reliability estimate for each:

```
trueAssociation <- 0.7
reliabilityOfMeasure <- 0.9

attenuationCorrelation(
  trueAssociation = trueAssociation,
  reliabilityOfPredictor = reliabilityOfMeasure,
  reliabilityOfCriterion = reliabilityOfMeasure)
```

```
[1] 0.63
```

3.9.2 Disattenuation of Observed Test–Retest Reliability Due to Measurement Error

Then, to disattenuate the observed test–retest reliability due to random measurement error, we can rearrange the terms in the formula above to estimate what the true stability of the construct is, according to Equation (3.22):

$$\begin{aligned} r_{x_{1_t} x_{2_t}} &= \frac{r_{x_1 x_2}}{r_{xx}} \\ \text{true correlation of construct across time points} &= \frac{\text{test–retest correlation of measure}}{\text{reliability of measure}} \end{aligned} \tag{3.22}$$

Find the true correlation of the construct across two time points given an observed correlation if the observed reliability of the measure is .9. The `petersenlab` package (Petersen, 2024) contains the `disattenuationCorrelation()` function that estimates the observed association given the true association and the reliability of the predictor and criterion. When dealing with test–retest reliability, the predictor and criterion are the same, so we use the same reliability estimate for each:

```
observedAssociation <- as.numeric(cor.test(
  x = mydata$time2,
  y = mydata$time3)$estimate)

disattenuationCorrelation(
  observedAssociation = observedAssociation,
  reliabilityOfPredictor = reliabilityOfMeasure,
  reliabilityOfCriterion = reliabilityOfMeasure)
```

```
[1] 0.9362
```

The next section discusses method biases, including what they are, why they are important, and how to account for them.

3.10 Method Bias

3.10.1 What Method Biases Are and Why They Are Important

Method bias is the influence of measurement on a person's score that is not due to the person's level on the construct. Method bias is a form of systematic, nonrandom measurement error because it involves errors/biases that are shared/common to/correlated for two measures.

Method bias is problematic because it can bias estimates of a measure's reliability and validity, and its association with other measures/constructs. There is a biasing effect of assessing two or more constructs with the same method on estimates of the association between them—that is, some of the observed covariance may be due to the fact that they share the same method of measurement.

Although estimates vary widely from study to study, construct to construct, and measure to measure, estimates of method bias tend to be substantial. Meta-analyses suggest that around 20–40% of a given measure's variance reflects method variance (for a review, see Podsakoff et al., 2012). According to Podsakoff et al. (2012), response styles inflate the magnitude of observed correlations by 54% if there is a positive correlation between the constructs and a positive correlation between the response styles. Response styles inflate the magnitude of observed correlations by even more (67%) if there is a negative correlation between the constructs and a negative correlation between the response styles. Moreover, the method biases can also weaken the association between measures. If the direction of correlation between the constructs (e.g., positive) is incongruent with the direction of the correlation between the response styles (e.g., negative), the association between the measures is attenuated.

3.10.2 Types of Method Biases

There are many different types of method biases, including the effects of

- same source: Whether ratings or information comes from the same source can lead to systematic method variance.
- response styles (response biases): Response styles are systematic ways of responding to items or questions that are not due to the construct of interest. Different types of response styles can lead to systematic method variance across constructs. Examples of response styles include
  - acquiescence: the tendency to agree with items (both positively and negatively worded), regardless of content
  - disacquiescence: the tendency to disagree with items (both positively and negatively worded), regardless of content
  - extreme: the tendency to endorse items on the extremes (positive or negative) of the response scale, regardless of content
  - midpoint/central tendency: the tendency to endorse items on the middle scale category, regardless of content
  - noncontingent: the tendency to respond to items carelessly, randomly, or nonpurposefully
  - social desirability: the tendency to respond to items in a way that others will believe the respondent is better than they actually are. Social desirability includes multiple components, including self-deception and impression management. Impression management is the conscious effort to control the way people view oneself or the way one presents oneself.
- proximity: Items (of unrelated constructs) that are in closer proximity to one another in a measure are more likely to be correlated than items that are further apart. The associations of items assessing unrelated constructs are weakest when the pair of items are six or more items apart from each other. There is an exception for reversed items. For reversed items (e.g., reverse scored), the further apart the reversed item is placed from another item, the higher the negative correlation.
- item wording: Whether items are positively or negatively worded can influence the association between items (i.e., like-worded items go with like).
- item context: Manipulating the context of a measure can have an effect on the perceived association. You can observe a different association between constructs when you put the measure in a negative context versus in a positive context.
- item ambiguity: When respondents are less certain of how to accurately answer a question, it increases the likelihood that they will rely on response styles or succumb to the effect of context when responding, which biases scores.
- motivational factors: Researchers need to motivate participants to put in the cognitive effort to answer questions correctly, to decrease the likelihood that they will succumb to responding based on response style only. Also, researchers should decrease pressures for "optimal responding" on respondents. That is, researchers should attempt to reduce social desirability effects.
- the ability of the respondent: Respondents with low cognitive abilities, less education, more fatigue (after lots of questionnaires), less experience with the topic being assessed, and more uncertainty about how to respond are more likely to succumb to biasing factors.

3.10.3 Potential Remedies

There are numerous potential remedies to help reduce the effects of method bias, including

- obtain measures of the predictor and criterion from multiple or different sources
- create temporal, physical, and/or psychological separation between the predictor and criterion
- eliminate common scale properties; vary the scale types and anchor labels
- remove ambiguity in items—keep questions simple, (behaviorally) specific, and concise
- reduce social desirability in item wording
- provide anonymity to respondents
- for estimating the prevalence of a sensitive behavior, use the randomized response model (S. J. Clark & Desharnais, 1998), as described below
- provide incentives (e.g., compensation) to respondents
- reverse score some items to balance positively and negatively worded (i.e., reverse-scored) items
- separate items on the questionnaire to eliminate proximity effects
- maximize respondent motivation to give an accurate response
- develop a good cover story and clear instructions, with reminders as necessary
- match the measure difficulty to the respondent's ability; minimize frustration
- use confirmatory factor analysis or structural equation modeling, as described in Section 7.10.1
- use mixed models, as described below

3.10.3.1 Randomized Response Model

The randomized response model (S. J. Clark & Desharnais, 1998) is a technique to increase people's honesty in responding, especially around sensitive topics. The randomized response model sought to answer the question of how much information people withhold on surveys. It is no surprise that people lie on surveys, especially if the survey asks questions about sensitive topics, such as suicidality, criminal behavior (e.g., drug use), and child abuse or neglect. The randomized response model is based on the premise that respondents would give more honest answers if they could be certain that their survey answers would not reveal any sensitive information about themselves, and it is designed to reduce evasive-answer bias by guaranteeing privacy.

The randomized response model works by exploiting statistical probabilities. Consider a survey that asks respondents whether they have used various illegal drugs (e.g., "Have you used heroin?" or "Have you used cocaine?"). Each participant is given a coin. They are told to flip the coin before they answer each question. If the coin flip is "heads," they answer the question truthfully. If the coin flip is "tails," they answer the questions with a "yes" (i.e., the more sensitive answer) no matter what. That way, a yes could be due to the coin flip or could be the truth—so respondents feel as if they are divulging less.

Then, researchers can estimate the actual "no" rate by doubling the observed "no" rate for the sample. For instance, if 20% of people say they have never gambled and 80% say they have gambled, the actual "no" rate of gambling is 40%. That is, the actual gambling rate is 60%.

Findings using the randomized response model indicate that there are a lot of things that people are not sharing. The randomized response model is useful for estimating the true prevalence rate of sensitive behaviors. However, it is not great for individual differences research because you do not know whether any given individual truly engages in a behavior (because a "yes" response may have occurred merely because the coin flip was "heads").

Thus, the randomized response model is only helpful for knowing "true" overall rate of a given behavior in the sample.

3.10.3.2 Mixed Models

One approach to handle method biases is in mixed models, which are also called mixed-effects models, multilevel models, and hierarchical linear models (HLMs). Mixed models are similar to multiple regression in that they allow fitting a model that examines multiple predictor variables in relation to one outcome variable. One of the assumptions of multiple regression is that the data observations are independent from each other, as evidenced by the residuals being uncorrelated with each other. Mixed models are designed to handle nonindependent observations, that is, data that were collected from nonindependent units. Thus, mixed models can be fit to data that do not meet the nonindependence assumption of multiple regression.

Examples of data from nonindependent units include nested data, multiple membership data, and cross-classified data.

One example of nonindependent data is if your data are *nested*. When data are collected from multiple lower-level units (e.g., people) in an upper-level unit (e.g., married couple, family, classroom, school, neighborhood), the data from the lower-level units are considered nested within the upper-level unit. For example, multiple participants in your sample may come from the same classroom. Data from multiple people can be nested within the same family, classroom, school, etc. Longitudinal data can also be considered nested data, in which time points are nested within the participant (i.e., the same participant provides an observation across multiple time points). And if you have multiple informants of a child's behavior (e.g., parent-, teacher-, friend-, and self-report), the ratings could also be nested within the participant (i.e., there are ratings from multiple informants of the same child).

Another form of nonindependent data is when data involve *multiple membership*. Data involve multiple membership when the lower-level units belong to more than one upper-level unit simultaneously. As an example of multiple membership, children may have more than one teacher, and therefore, in a modeling sense, children "belong" to more than one teacher cluster.

Another form of nonindependent data is when data are *cross-classified* (also called crossed). Data are cross-classified when the lower-level units are classified by two or more upper-level units, but the upper-level units are not hierarchical or nested within one another. For instance, children may be nested within the crossing of schools and neighborhoods. That is, children are nested within schools, and children are also nested within neighborhoods; however, children attending the same school may not necessarily be from the same neighborhood. The data would still have a two-level hierarchy, but the data would be nested in specific school-neighborhood combinations. That is, children are nested within the cross-classification of schools and neighborhoods.

The general form of multiple regression is (see Equation (3.23)):

$$y = b_0 + b_1 x + e \tag{3.23}$$

where b_0 is the intercept, e is the error term, and b_1 is the slope of the predictor, x, that states the relation between the predictor (x) and the outcome (y). However, the slope can exist at multiple levels when data are nested within groups or when time points are nested within people. For example, you can observe different associations of a predictor with an outcome at the between-individual level compared to the within-individual level. That is, the association between the predictor and outcome can differ when comparing across individuals versus when looking at the association between the predictor and outcome based on fluctuations within the individual. When the associations differ at multiple levels, this is known as *Simpson's paradox*. For an example of Simpson's paradox, see Figure 3.24 (R. Kievit et al., 2013). Thus, it is important to consider that the association between variables may differ at different levels (person level, group level, population level, etc.).

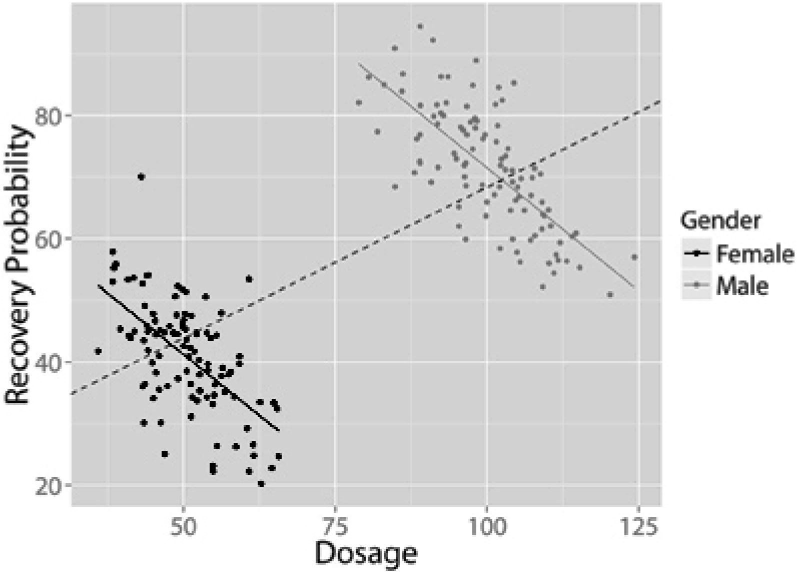


FIGURE 3.24 Example of Simpson's paradox, where the association between the predictor and outcome differs at different levels. In this case, the association between dosage and recovery probability is positive at the population-level (upper-level unit), but the association is negative among men and women separately (lower-level unit). (Figure reprinted from R. Kievit et al. (2013), figure 1, p. 2. Kievit, R., Frankenhuis, W., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology*, *4*(513). https://doi.org/10.3389/fpsyg.2013.00513)

To account for method biases in a mixed model, you could use scores from multiple measures and methods as the outcome variable. For instance, your mixed model could examine the association between income and optimism, where optimism is assessed by self-report, parent-report, and a performance-based measure. In the model, you could include, as a *fixed effect*, a factor that indicates the method of the measure, in addition to its interaction with income, to account for systematic ways in which the measurement method influences scores on the measures across the participants as a whole. You could also include a *random effect* of measurement method if you wanted to allow the effect of measurement method on the scores to differ from person to person.

3.11 Generalizability Theory

Up to this point, we have discussed reliability from the perspective of CTT. However, as we discussed, CTT makes several assumptions that are unrealistic (e.g., that all error is random). There are other measurement theories that conceptualize reliability differently than the way that CTT conceptualizes reliability. One such measurement theory is generalizability theory (Brennan, 1992), also known as G-theory and domain sampling theory. G-theory is described in detail in Chapter 5 on generalizability theory and is also discussed in the chapters on validity (Chapter 4, Section 4.7) and structural equation modeling (Chapter 7, Section 7.11).

3.12 Item Response Theory

In addition to CTT and G-theory, item response theory (IRT) is another measurement theory. IRT also estimates reliability in a different way compared to CTT, as described in Section 8.1.6. We discuss IRT in Chapter 8.

3.13 The Problem of Low Reliability

Using measures that have low reliability can lead to several problems, including

- over- or underestimation of the effects of interest. When sample sizes are large, measurement error tends to weaken effect sizes. However, when sample sizes are small to modest, measurement error can actually lead to overestimation of effect sizes as well as inaccurate estimates of the sign of the effect (positive or negative), due to the statistical significance filter (Loken & Gelman, 2017), as described in the section on "Assessment and the Replication Crisis" in the Introduction.
- reduced statistical power to detect the effects of interest
- failure to replicate findings
- invalid measurement for the proposed uses. As we described in the chapter on validity in Section 4.5, reliability constrains validity.

These concerns are especially important for individual differences research (e.g., correlational designs) or when making decisions about individual people based on their scores on measure. However, these concerns are also relevant for experimental designs.

3.14 Ways to Increase Reliability

Here are several potential ways to increase reliability of measurement:

1. Improve the clarity and homogeneity of items.
2. Use a structured response set and scoring system rather than unstructured ones.

3. Make sure participants know how to complete the task or items. For example, make sure they have appropriate instructions that are easily understood.
4. Adhere closely to the standardized procedures for administering and scoring a measure.
5. Make scoring rules as explicit as possible.
6. Train raters to some criterion of inter-rater reliability. Continue to monitor inter-rater reliability over time to reduce coder drift.
7. Use a more precise response scale. The smaller the number of response options, the greater the attenuation of a correlation due to measurement error (Rigdon, 2010).
8. Have a sample with a full range of variation in the items and measure.
9. If possible, do not use a difference score. This is discussed in greater detail in Section 3.5.8 and in Chapter 23 (Section 23.3.4.3.1) on cognitive assessment.
10. Remove items that reduce reliability. This may include items that have low variability because they are too easy or too difficult (i.e., items that show ceiling or floor effects), in addition to items that have a low item–total correlation.
11. Collect more information across people and time.
12. Aggregate: Add items, measures, observers, raters, and time points.
    a. Having more items that assess the same construct in the measure tends to increase reliability unless the additional items cause fatigue because the instrument is too long. Aggregation of like things reduces measurement error, because averaging together more items helps cancel out random error.
    b. Having multiple measures that assess the same construct can be beneficial, especially measures from different methods that do not share the same systematic errors (e.g., observation in addition to parent report).
    c. Items that aggregate a large amount of information tend to have strong predictive validity. In general, for predictive accuracy, it is often better to assess more things less well than to assess one thing extraordinarily well. Aggregation is a double-edged sword: It is your friend as a researcher, because of increased predictive utility, but it is your enemy as an interpreter because it is hard to interpret the theoretical meaning of predictive associations based on an aggregated variable because the measure is more general and nonspecific. If aggregation increases error, test–retest reliability will decrease. If error stays the same when aggregating, test–retest reliability will increase.
13. Use a latent variable approach to aggregate the multiple items, measures, observers, raters, or time points. Latent variable approaches include structural equation modeling and item response theory. Estimating a latent variable allows identifying the variance that is common to all of the measures (i.e., a better approximate of true score variance), and it discards variance that is not shared by all measures (i.e., error variance). Thus, latent variables have the potential to get purer estimates of the constructs by reducing measurement error.

3.15 Conclusion

Reliability is how much repeatability, consistency, and precision a measure's scores have. Reliability is not one thing. There are multiple aspects of reliability, and the extent to which a given aspect of reliability is important depends on the construct of interest. In general, we want our measures to show strong inter-rater, intra-rater, immediate parallel-forms, and internal consistency reliability. However, the extent to which our measures should show test–retest and delayed parallel-forms reliability depends on the stability of the construct and the time lag. Moreover, although we want our measures to show strong internal consistency reliability, the internal consistency reliability of a measure can be *too* high, and for some constructs (i.e., formative constructs), internal consistency would not be expected. In addition, reliability is not a characteristic that resides in a test. The reliability of a measure's scores reflects an interaction of the properties of the test with the population for whom it is designed and the sample, situation, and context in which it is administered. Thus, when reporting reliability in papers, it is important to adequately describe the aspects of reliability that have been considered and the population, sample, and context in which the measure is assessed. However, it is not enough for measures to be consistent or reliable; our interpretation of the measures' scores should also be accurate, meaningful, and useful for the intended uses. In other words, our interpretation of the measures' scores should be valid for the intended uses. In the next chapter (Chapter 4), we discuss validity of measurement.

3.16 Suggested Readings

For more information on inter-rater reliability, we suggest reading Gwet (2021a) and Gwet (2021b), for categorical and continuous ratings, respectively.